9-2011

Binary ROCs and Their Implications for the Measurement of Memory

Chad Dube
University of Massachusetts Amherst, chaddube@gmail.com

BINARY ROCS AND THEIR IMPLICATIONS FOR THE MEASUREMENT OF MEMORY

A Dissertation Presented

by

CHAD M. DUBE

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2011

Psychology

BINARY ROCS AND THEIR IMPLICATIONS FOR THE MEASUREMENT OF MEMORY

A Dissertation Presented

by

Chad M. Dube

Approved as to style and content by:

_____
Caren M. Rotello, Chair

_____
Jeffrey J. Starns, Member

_____
Rebecca E. Ready, Member

_____
Donald L. Fisher, Member

_____
Melinda A. Novak, Department Head
Department of Psychology

DEDICATION

To my Mother and Father.

ACKNOWLEDGEMENTS

My proudest accomplishments in graduate school are a direct result of having followed Caren Rotello's advice. But she doesn't simply provide direction. She can impart the kind of humor and evenness that brightens up even the most difficult situations. She also sets an example that in terms of rigor and devotion to science cannot be surpassed, and this is something that has registered deeply with her students and with some of the most respected people in her field. So thanks, Caren, for teaching me so much and for being such a great person to work with.

Profuse thanks are due also to my committee members Don Fisher, Jeff Starns, and Becky Ready, who have taken time out of their busy schedules to help me get to this point. A special thanks to Jeff for allowing me to distract him from his work when I wanted to learn about using optim() in R.

Thanks to Neil Macmillan, Dennis Delprato, and Chuck Clifton for their assistance, advice, and words of encouragement. Thanks to Evan Heit for the knowledge he has shared and for having the rare skill of successfully conveying humor in email messages.

I have been lucky to have worked in the lab with such excellent people. Thanks to Aycan Kapucu, Min Zeng, Angela Pazzaglia, and Davide Bruno. Thanks also to our research assistants: Alyssa Champagne, Bethany Eaton, Jessica Regan, Laura McDonough, and Samantha Westray.

A special thanks to Ben Zobel for E Prime help, but mostly for being a great friend.

Finally, thanks to Lindsay Demers. I don't even know where to begin. When this defense business is over with, I would like to take you out for dinner and margaritas.

ABSTRACT

BINARY ROCS AND THEIR IMPLICATIONS FOR THE MEASUREMENT OF MEMORY

SEPTEMBER 2011

CHAD DUBE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Caren M. Rotello

Bröder and Schütz (2009) have argued that the curvature typically observed in recognition memory receiver-operating characteristics (ROCs) is a by-product of the ratings task often used to obtain them. According to those authors, ROCs collected by experimentally manipulating response bias are linear and consistent with the assumptions of threshold and multinomial processing tree (MPT) models. Two experiments are reported which are broadly consistent with previous work by Dube and Rotello (under review) in showing that ROCs are curved and consistent with signal detection theory (SDT) regardless of the procedure used to obtain them. These results have implications for how accuracy is measured in tasks requiring binary responses. It is suggested that the use of statistics consistent with the assumptions of threshold models (e.g. percent correct, hits minus false alarms) should be avoided, as they are likely to produce statistical errors in several areas of the literature where ROCs have been found to be curved (Rotello, Masson, & Verde, 2008; Dube, Rotello, & Heit, 2010; 2011). SDT-based measures and ROC analysis are recommended to complement or replace analyses based on threshold statistics.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Overview

A fundamental attribute of the human memory system is the capacity to recognize. The experimental study of recognition memory is often conducted via an item recognition technique. In a prototypical item recognition task, participants are presented with a list of words to study followed by a test list that contains a mixture of previously studied words (old items or 'targets') as well as words that were not on the study list (new items or 'lures'). The participant's task is to indicate whether a given test item is old or new. This procedure yields four response types: correct 'old' responses (called 'hits'), correct 'new' responses ('correct rejections'), incorrect 'old' responses ('false alarms'), and incorrect 'new' responses ('misses'). Since the hit and miss rates necessarily sum to one, as do false alarm and correct rejection rates, researchers typically focus on hits and false alarms.

Accuracy in the recognition task is called 'sensitivity.' It refers to the participant's ability to discriminate between old and new items on the test. One fairly intuitive way to measure sensitivity is to simply consider the proportion of old items correctly called 'old.' Unfortunately, this proportion is influence by response bias, the participant's overall willingness to say 'old' or 'new' at test. For example, a participant adopting a very liberal response bias might say 'old' on almost every trial, which could produce a hit rate near one even if true sensitivity is actually very low. The influence of response bias on the hit and false alarm rates makes the assessment of true sensitivity

difficult. For this reason, quantitative models are often used to provide measures of sensitivity and bias.

Some of the more intuitive accuracy measures that attempt to correct or control for bias are situated in a threshold framework (Egan, 1958; Krantz, 1969). For instance, one type of threshold model, called the double high-threshold model, predicts that participants who are unbiased will produce a false alarm rate (*F*) equal to zero and a hit rate (*H*) equal to true sensitivity. As the bias to say 'old' increases, *F* and *H* are expected to increase by the same amount. This prediction follows from the processing assumptions of the threshold model, in which participants are assumed to respond on the basis of a small number of discrete mental states. Under this assumption, simply subtracting *F* from *H* should provide a good estimate of sensitivity. *H*-*F* has been advocated for recognition data by Snodgrass and Corwin (1988), who called the measure P_r.

Another class of models for measuring recognition is situated in a signal detection framework (Green & Swets, 1966; Macmillan & Creelman, 2005). Signal detection theory (SDT) generally predicts that as the 'old' bias increases, *H* and *F* will always increase, but the rate at which they increase will not be constant as is assumed by threshold models. This follows from the processing assumptions made by SDT models: participants are assumed to respond on the basis of a continuously-distributed memory strength variable rather than the discrete states assumed in threshold models. The most basic signal detection model (called the 'equal variance' model for reasons to be elaborated on below) assumes a different measure for accuracy in the recognition task, called *d'*.

Importantly, the validity of the measurement indices provided by threshold and signal detection models can have major implications for the conclusions researchers reach in their analyses. For instance, Rotello, Masson, and Verde (2008) conducted simulations to estimate the probability of falsely concluding that two groups differ in accuracy (when only bias differs) in two basic conditions. In one condition, the assumptions of a threshold model were met (the data were generated using rectangular distributions); in the other condition, the assumptions of a signal detection model were met (the data were generated using Gaussian distributions). For each condition, the parameters of both models were used to measure accuracy in the simulated data. The results showed that Type I error rates remained near the nominal alpha level of .05 when a given measure was consistent with the generating model. They also showed that the probability of a Type I error when the wrong measure was applied was drastically high (near .9 in several instances) and, worse yet, increases in the number of observations actually increased the error rate. These data suggest that when the assumptions of a given model are not valid, the use of that model's parameters may lead to persistent Type I error with a very high probability. The outcome predicted by Rotello et al. (2008) has recently been confirmed empirically by Dube, Rotello, and Heit (2010), who showed that decades of theoretical work in deductive reasoning were founded on an effect in *H*-*F* that was in fact a Type I error of exactly this sort.

Taken together, these results carry implications for previous studies of recognition, source monitoring, reasoning, and even social cognition, where threshold statistics have often been applied in the absence of any examination of the assumptions inherent in those statistics (e.g. Bayen, Murnane, & Erdfelder, 1996; Bröder & Meiser,

2007; Klauer, Musch, & Naumer, 2000; Sherman, 2006). If the assumptions of those measures are not valid, then the conclusions that were based on those measures are also likely to be invalid, as they have been in previous studies of deductive reasoning.

Fortunately, researchers can avoid elevated Type I error rates by evaluating the assumptions of threshold and signal detection models, which provide many of the accuracy measures commonly used in recognition. This can be done by directly examining the way in which changes in response bias alone affect *H* and *F*. That is, one can plot *H* against *F* as a function of response bias at a given level of accuracy. These plots are called *isosensitivity curves* or, more commonly, *receiver-operating characteristics* (ROCs). Crucially, ROCs generated with simple threshold models are linear while those generated with simple signal detection models are curved. As noted in a recent review by Wixted (2007), recognition ROCs have nearly always been found to be curved and therefore inconsistent with basic threshold models. This is consistent with conclusions reached in the perception literature as well, which is likely a major reason why signal detection models have been widely adopted for perception and recognition (Green & Swets, 1966; Macmillan & Creelman, 2005; Swets, 1986a; 1986b).

Classic ROC-based arguments in favor of SDT have recently been challenged by Bröder and Schütz (2009). The authors noted that much of the empirical ROC data that have been taken to validate SDT analyses were collected using a confidence ratings task to produce the different levels of response bias. This is true of countless studies conducted in a variety of literatures over the past 50 years, of which the reasoning analyses reported by Dube et al. (2010) are one example. Following a previous demonstration by Malmberg (2002), Bröder and Schütz argued that variability in how

participants use confidence ratings can produce curvature even if the underlying process is best described by a threshold model. As I will detail below, threshold models do clearly predict linear ROCs when the data are collected using an experimental manipulation of response bias (rather than ratings). Thus, the key question is whether these Yes/No (or binary) ROCs are linear. If it can be concluded that these functions are linear, then a large number of ratings results such as the ones reported by Dube et al. will be called into question, as will the innumerable studies that have used SDT statistics such as d'. Bröder and Schütz reported a meta-analysis and three new experiments that allowed comparisons of SDT and threshold models using binary ROCs, and concluded that the data "…do not speak against threshold models. Rather, the 2HTM (double high-threshold model) tends to fit data better than SDT in bias manipulation experiments" (p. 600).

In what follows, I will describe the study by Bröder and Schütz (2009) and the shortcomings of that study as pointed out by Dube and Rotello (under review). I will show that, although the claims made by Bröder and Schütz were challenged by Dube and Rotello (submitted), there are limitations to the latter study that may confine Dube and Rotello's conclusions to the particular tasks and/or materials they used. Following this, I will present data from two new experiments that attempt to address these issues. First, however, I will present a more thorough discussion of the threshold and signal detection models, their assumptions, and related predictions.

Threshold Theory and Signal Detection Theory

Threshold theory (Egan, 1958; Krantz, 1969), broadly speaking, assumes that recognizing is an all-or-none affair: old items at test must be strong enough to pass a threshold for remembering. The 'strength' of an item in threshold models, then, is important only insofar as it either meets or fails to meet a value required to pass the threshold. Although this idea is not widely supported in perception and recognition (Macmillan & Creelman, 2005; Swets, 1986a; 1986b; Wixted, 2007), threshold theories are still popular in many areas of cognitive science, where they are frequently implemented in the form of *multinomial processing tree* (or MPT) models (e.g. Batchelder & Riefer, 1990; Batchelder & Riefer, 1999; Bayen, Murnane, & Erdfelder, 1996; Bröder & Meiser, 2007; Klauer, Musch, & Naumer, 2000).

The basic assumptions of MPT models can be examined using the double-high threshold model (2HTM) depicted in the left panel of Figure 1. In this model, old item trials result in a 'detect' state with probability p_o. When participants enter this mental state, they are assumed to always respond 'old.' With probability $1\text{-}p_o$, however, participants fail to detect the status of old items (i.e. the old items do not pass the threshold) and enter a non-detect state. In the non-detect state, participants are assumed to have no information about the item and must respond on the basis of random or biased guessing, reflected in the parameter b. Similar logic applies to lure trials: participants detect that new items are new with probability p_n, fail to detect that they are new with probability $1\text{-}p_n$, and guess according to b. Hit and false alarm rates can be estimated by summing the probabilities for all branches that begin with either old or new items and end with an 'old' response. Thus, $H = p_o + (1\text{-}p_o)b$ and $F = (1\text{-}p_n)b$. If the probability of

correctly detecting an item's status can be assumed to be independent of item status, i.e. $p_o = p_n = p$, a simple measure of sensitivity can be obtained by subtracting hits and false alarms: $H - F = p$. Thus, the intuitive measure $H - F$ actually implies (or, is a parameter of) a particular version of the double high-threshold model.

The three states assumed by the 2HTM have implications for the form of the binary ROC. Specifically, the ROC is predicted to be linear with a y-intercept equal to p_o and upper-x intercept equal to 1-p_n. Different (F, H) pairs on the same ROC reflect changes in response bias (here, changes in b) only. An ROC generated with the 2HTM is illustrated in the right panel of Figure 1. Empirical ROCs can be used to evaluate the assumptions of the 2HTM, and thus the validity of its measurement indices. If the observed binary ROC is linear, for example, but the slope of the line is not equal to one ($p_n \neq p_o$), then one cannot estimate sensitivity using H-F because the statistic will no longer be independent of response bias. If the observed ROC is not linear, this implies a more fundamental assumption of the 2HTM is invalid: participants cannot be assumed to be responding on the basis of the three mental states depicted in Figure 1. In this case, changes in response bias only are likely to produce a Type I error in threshold parameters such as percent correct, H-F, or p_o (Dube, Rotello, & Heit, 2010; 2011; Rotello, Masson, & Verde, 2008).

Curvature in empirical ROCs is broadly consistent with the fundamental assumptions of signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005). Unlike the 2HTM, SDT assumes that the specific strength values of individual items do matter as there is no fixed threshold for old or new item recognition. That is, participants are assumed to operate on the basis of Gaussian distributions of

memory strength. As illustrated in the left panel of Figure 2, the basic SDT model assumes that there are two such distributions, one each for old and new items, with higher mean strength assigned to old items as a consequence of their recent presentation on the study list. The means of the old and new item distributions are estimated by μ_o and μ_n, respectively, and the corresponding standard deviations are σ_o and σ_n.[1] To make a given recognition decision, participants are assumed to compare the strength of a given test item to an old/new criterion, c_x. If the item is strong enough to exceed the criterion, the participant responds 'old', otherwise he or she responds 'new.' The hit rate, then, is the area under the old item distribution that falls to the right of the criterion, and the false alarm rate is the area under the new item distribution that falls to the right of the criterion. Miss and correct rejection rates correspond to the areas under the corresponding distributions that fall to the left of the criterion. Importantly, the model depicted in Figure 2 assumes that the variance of the old item distribution is equal to that of the new item distribution. If this assumption holds, then sensitivity in the SDT model corresponds to the difference in mean activation divided by the common standard deviation. This measure, called d', can be estimated by applying an inverse Gaussian transform to H and F.

$$d' = z(H) - z(F) \qquad (1)$$

Analogous to the case of the threshold model, however, use of d' when the equal-variance assumption is not met can greatly elevate the risk of falsely concluding that two conditions differ in accuracy when they differ only in bias (Rotello et al., 2008).

[1] The mean and standard deviation of the lure distribution are typically set to 0 and 1, without loss of generality.

As for threshold models, the assumptions of the equal-variance SDT model can be evaluated using ROCs. To generate a predicted binary ROC with the SDT model, a single value of d' is chosen and the value of the criterion c_x is varied across several levels. At each level of c_x, the areas under the old and new item distributions to the right of the criterion are plotted in ROC space. When performance is above chance, this method produces a curved ROC like the one in the right panel of Figure 2. Conservative response biases in SDT imply a rightward criterion and produce operating points that fall near the origin, where H and F are relatively low. As the criterion moves leftward, responding becomes more liberal and both H and F increase. This produces points that fall nearer to the upper right corner of the ROC plot. As can be seen in Figure 2, the ROC curve predicted by the equal-variance model is symmetrical about the minor diagonal. When the ROC is plotted on z coordinates, this implies a straight line with unit slope (estimated in SDT as the ratio σ_n/σ_o). If the observed ROC is not consistent with the slope prediction (implying an asymmetric ROC and zROC with nonunit slope), then the equal-variance assumption is not warranted. In this case, one can adopt an unequal-variance SDT model that allows σ_o and σ_n to vary. As zROC slopes are often found to be less than 1 in the recognition literature, the unequal-variance SDT model is usually preferred over the equal-variance model (Glanzer, Kim, Hilford, & Adams, 199; Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, McKoon, & Tindall, 1994).

An appropriate measure of sensitivity in the unequal-variance model expresses the difference between the means of the target and lure distributions in terms of the root-mean-square of the target and lure standard deviations. This measure, called d_a, can be obtained given H, F, and an estimate of zROC slope. As for H-F, d_a can also be defined

using the parameters of its implied model. d_a is defined below for calculation given H, F, and a slope estimate. The measure implies an unequal-variance model with $\sigma_n = 1$ and $s = 1/\sigma_o$, the slope of the zROC (see Macmillan & Creelman, 2005, for derivation).

$$d_a = \sqrt{\frac{2}{1+s^2}}\left[z(H) - sz(F)\right] \qquad (2)$$

If the observed ROC is linear, however, it implies that the basic processing assumptions of the SDT model are in error. In this case, an appropriate threshold model might be adopted in order to minimize the probability of committing a Type I error.

In the recognition literature, the ROC predictions of the SDT model have most frequently been evaluated using ratings-based (rather than binary) ROCs. In the ratings method, participants are asked to follow each Old/New decision with a confidence rating, e.g. on a scale of 1-6 ranging from 'Sure Old' to 'Sure New.' The ratings ROC is constructed by treating these decisions as Old/New responses made with different response biases, and cumulating H and F across the levels 1-6. By this interpretation, a rating of 1 would correspond to the most conservative response rule in the task and would produce an operating point near the origin in Figure 2. The second point on the ratings ROC is obtained by adding H and F at a rating of 2 to the corresponding proportions at a rating of 1. The ratings-contingent proportions are cumulated in this way until all responses have been summed, producing a 6-point ROC with the rightmost point necessarily falling at (1,1). The data can then be modeled using SDT by assuming that participants maintain several criteria simultaneously. Using 5 criteria (including the old/new criterion in Figure 2) participants can partition the strength axis into the 6 response categories. Here, the areas under the old and new distributions to the right of

the rightmost criterion would correspond to *H* and *F* conditional on a rating of 1. The area between the two rightmost criteria would correspond to *H* and *F* at a rating of 2, and so on. A predicted ROC can then be plotted by cumulating the areas across the 6 partitions. As ratings ROCs have virtually always been found to be curvilinear and consistent with the assumptions of the unequal-variance SDT model, researchers have embraced that model for analyses of recognition data (Macmillan & Creelman, 2005; Swets, 1986a; 1986b; Wixted, 2007).

Unfortunately, though threshold models such as the 2HTM clearly predict linear binary ROCs, it is not the case that they necessarily predict linear ratings ROCs. As demonstrated by Malmberg (2002), the exact form of the ratings ROC predicted by the 2HTM depends on how the states are mapped onto levels of confidence. For instance, one intuitive way to handle ratings data is to assume that the detect states always result in the highest confidence ratings corresponding to those states ('sure old' or 'sure new'). The non-detect state, however, could result in any level of confidence. When this mapping is used, the threshold model predicts linear ROCs similar to those predicted for binary data. When the detect states are allowed to map onto any level of confidence, however, the 2HTM generates ROCs that look similar to the curved functions generated by SDT. In other words, assuming there is inter- and/or intra-participant variability in how ratings are used, a curved ratings ROC could be observed empirically even if participants are actually responding on the basis of a threshold process.

In recent work, Klauer and Kellen (2010; 2011) have adopted this last assumption and proposed an extension of the basic 2HTM that incorporates variability in scale usage. In their model (depicted in Figure 3), participants responding on the basis of a given

detect state distribute their responses across the high-, middle-, and low-confidence categories (for the 6-point rating) according to the parameters s_h, s_m, and s_l. In this way, the model can produce curved ROCs while still assuming that processing is thresholded in nature. Importantly, the validity of the claim that curvature in ratings ROCs is due to scale-usage variability rests entirely on one recent study conducted by Bröder and Schütz (2009). We will now turn to a discussion of that study and the issues related to it.

Are Ratings ROCs Artifactually Curved?

The possibility of artifactual curvature in ratings ROCs would have major implications for past research in recognition that has adopted an SDT analysis. If ratings ROCs are curved solely as a result of the procedure, then past conclusions regarding accuracy on recognition tests could reflect Type I error due to the use of inappropriate measures (e.g. d', d_a; Dube et al., 2010; Rotello et al., 2008). The key question, then, is whether binary ROCs are typically curved or linear. The construction of binary ROCs can be accomplished by manipulating response bias experimentally. Manipulations of this sort include conditions that differ in the actual or perceived proportion of test items that are old, or rewards and penalties associated with different kinds of responses. These manipulations typically require either a large number of participants or multiple sessions per participant in order to obtain stable operating points across the conditions. Perhaps for this reason, recognition studies have typically involved the ratings method. Bröder

and Schütz (2009) managed to locate 81[2] binary datasets, however, and used them to compare the fit of the 2HT and unequal-variance SDT models.

Bröder and Schütz (2009) found that, contrary to previous results for ratings ROCs that clearly supported SDT (Swets, 1986a; 1986b), the models fit the data about equally well. They also reported three new experiments designed to compare the models using 5-point binary ROCs constructed from a base rate manipulation. Specifically, the percentage of test items that were old was varied across 5 levels from 10% to 90%, and participants were informed of this. In Experiments 1 and 2, bias was manipulated between-participants, but the experiments differed in whether the stimuli were words (Experiment 1) or pictures (Experiment 2). The results from these experiments were analyzed both at the group level, and at the level of three subgroups that differed in overall accuracy. The authors found that the 2HTM provided a good fit to the data, and a numerically better fit than SDT in terms of the likelihood statistic G^2, but that the SDT model was only rejected in the group analysis of Experiment 2. Experiment 3 was similar to Experiment 2 in that pictures were used, but in order to minimize the possibility of accuracy changes across conditions bias was manipulated within-participants. Contrary to the results from Experiment 2, the results for this experiment showed a good fit for both models, and the SDT model provided a numerically better fit to the data in the low accuracy subgroup. Bröder and Schütz concluded that ratings

[2] This number differs from the one published by Bröder and Schütz (2009). As of this writing, a revised analysis is being prepared as an erratum to the original paper (Arndt Bröder, personal communication). This writing will consider the data as they currently stand, which are nonetheless consistent with related conclusions reached in Dube and Rotello (under review).

ROCs are not useful for discriminating SDT and threshold processes, and that analyses of binary ROCs tend to support the 2HTM over the SDT model.

Dube and Rotello (under review) discussed a number of limitations of the meta-analysis and new experiments reported by Bröder and Schütz, and reported new data that contradict the conclusions reached by the latter authors. Considering first Bröder and Schütz' meta-analysis, Dube and Rotello noted that most of the data included in that analysis consisted of 2-point ROCs. Though Bröder and Schütz themselves commented on the fact that both models can perfectly fit 2-point ROCs, they were included nonetheless. Dube and Rotello reported that when the 2-point data are excluded, the remaining 19 datasets are clearly better fit by SDT both in comparative and absolute fits. The authors also reported a meta-analysis of 54 datasets from the perception literature, all of which consisted of ROCs with 3 or more points that were obtained for individual participants. 85% of these cases were better fit by SDT than the 2HTM, and the overall fit of the SDT model was superior to that of the 2HTM. Consistent with the better overall fit of the model, SDT was rejected only 3 times according to a compromise-power G^2 criterion (also used by Bröder and Schütz), while the 2HTM was rejected 11 times by the same criterion.

Regarding Bröder and Schütz' newer data, Dube and Rotello pointed out that the only experiment in which SDT was actually rejected (Experiment 2) used picture stimuli rather than words. As Onyper, Zhang, and Howard (2009) recently found that ratings ROCs for pictures were more linear than those for words, it is unclear whether the Bröder and Schütz result may actually indicate a processing difference for words and pictures consistent with previous research on picture recognition (e.g. Paivio, 1971). Dube and

Rotello also pointed out that bias can be manipulated by either misleading participants as to the proportion of old items at test, or actually varying the proportions and informing participants of this as did Bröder and Schütz. Unfortunately, there is evidence from perception and recognition studies that the latter manipulation may lead to distortions of the ROC data due to changes in accuracy and/or the underlying ROC slope across conditions (Balakrishnan, 1999; Markowitz & Swets, 1967; Mueller & Weidemann, 2008; Schulman & Greenberg, 1970; Treisman & Faulkner, 1984; Van Zandt, 2000). Related to this is another problem specific to the procedure used by Bröder and Schütz: the number of observations for either targets or lures is much smaller in the extreme conditions than the middle conditions (e.g. only 6 targets or lures in the 10% and 90% conditions of Experiment 1). This could decrease the stability of the resulting operating points relative to the other 3 conditions. Finally, Bröder and Schütz did not in any case report a true individual-participant analysis, which raises the question of whether their analyses may have been influenced by distortions from averaging (Estes, 1956; Pratte, Rouder, & Morey, 2010).

To address these concerns, Dube and Rotello used a within-participants manipulation of implied (rather than actual) base rates, used both word and picture stimuli, included direct comparisons of ratings and binary ROCs (Experiment 1), and conducted analyses at both the group- and individual-participant level for picture data (Experiment 2). Their results showed that a) ratings and binary ROCs for word stimuli were similar in form and consistent with SDT but not the 2HTM, and b) binary ROCs for picture stimuli were generally curved and consistent with SDT (but not the 2HTM), at both the individual and group levels. Dube and Rotello (under review) concluded that

binary and ratings ROCs are curved. This indicates that SDT analyses should be preferred over threshold analyses in order to minimize the risk of committing a Type I error, and is consistent with previous conclusions from recognition, perception, and reasoning studies to the same effect (Dube et al., 2010; Rotello et al., 2008).

While the claims made by Bröder and Schütz are clearly challenged by Dube and Rotello's new data, there are limitations to the latter study that may confine Dube and Rotello's conclusions to the particular tasks and/or materials they used. Specifically, a single bias manipulation was used, which was itself quite different from the one used by Bröder and Schütz. Another limitation of the study by Dube and Rotello is that most of the data were obtained using picture rather than word stimuli. As word stimuli have been quite commonly used in previous comparisons of threshold and SDT models in recognition, a closer examination of binary ROCs for word stimuli is needed. A related limitation is the absence of individual-participant data for words. For this reason, it is unclear whether distortions due to averaging across participants may have influenced the results for words (Estes, 1956; Pratte, Rouder, & Morey, 2010). The experiments to be reported below expand on the work reported by Dube and Rotello by including another manipulation of response bias, as well as the one originally used by Bröder and Schütz. For the latter, a manipulation was included to assess the degree to which an actual base rate manipulation may produce changes in accuracy across conditions. The new experiments also focused solely on word stimuli. In all cases, analyses were conducted at both the group and individual-participant levels in order to ensure against problems inherent in averaging. These new experiments, together with the previous work reported

by Dube and Rotello, allow a more comprehensive evaluation of the form of binary ROCs in recognition

CHAPTER II

EXPERIMENT 1

The goal of Experiment 1 was to generalize the results of Dube and Rotello (under review) to a different kind of bias manipulation. Here the method was a slight modification of a bias procedure previously used by Egan, Schulman, and Greenberg (1959) for tone detection. In this procedure, individuals participated in three binary sessions in each of which they were asked to maintain a 'lax', 'medium', or 'strict' criterion (within-participants). The precise instructions corresponding to the lax, medium, and strict conditions also varied across sessions, so that overall participants were exposed to 9 levels of the response bias variable. The criteria for a given test were defined for the participants by asking them to keep their hit rate, false alarm rate, or overall yes rate within bounds determined by the experimenter. Participants were given feedback indicating how well they were approximating the desired response rate for a given condition, at pre-determined intervals. Following Egan et al., the same participants also responded in a standard ratings experiment, allowing a comparison of binary and ratings ROCs. All participants studied and were tested on words.

Method

Participants

Twenty-seven psychology students at the University of Massachusetts participated; they received either course credit or payment in exchange for their participation. Several datasets were lost or excluded due to experimenter error (4),

programming error (3), failure to complete all of the sessions (2), or chance responding (4). The results reported below are based on the remaining 14 participants.

Design

The binary condition of Experiment 1 used a 2 (old or new test item) x 9 (criterion condition: strict, medium, or lax, in 3 clusters) repeated-measures design. Each individual participated in 5 half-hour sessions (1 practice session, 1 ratings session, and 3 binary sessions), each separated by a 1 to 2 day inter-session interval. In the first binary session, participants were randomly assigned to one of three bias clusters, in each of which they were instructed to adopt criteria that amounted to a 'strict' (S), 'medium' (M), or 'lax' (L) response bias. The cluster was sampled without replacement from the set of three clusters (see Table 1), and only one of the three clusters was used per session. The order of the criterion conditions for a given session was randomly sampled, without replacement, from the set {(LMS), (SLM), (MSL)}. Each session consisted of 3 study-test cycles corresponding to the 3 criterion conditions. Each study-test cycle consisted of 135 study and 150 test words (75 targets and 75 lures). Only a subset of 75 target items from a given study list was (randomly) presented on the test list, and participants were informed of this. This was done in order to render the current design more comparable to that of Experiment 2.

The ratings condition of Experiment 1 followed a procedure similar to that of the binary condition, but consisted of only a single session containing 3 study-test cycles. This session was administered as either the second or fifth session of Experiment 1, the order being counterbalanced across participants. In a given cycle, participants studied

135 words and were tested on 75 targets and 75 lures. In this condition there was no criterion manipulation: participants were instead asked to follow each old/new response with a confidence rating.

In order to help participants orient to the task, the initial session of Experiment 1 consisted of practice trials using both the binary and ratings methods described above. The practice session consisted of 3 study-test cycles, two of which used a test procedure randomly selected from the set of 9 binary conditions in Table 1 (a different procedure was used for each test), and one of which used the ratings procedure.

Stimuli

The stimulus pool used in Experiment 1 consisted of 3,150 singular nouns selected from the MRC psycholinguistic database (Coltheart, 1981). They were 5-8 letters long and had a written frequency < 200 (Kučera & Francis, 1967). For the binary condition, the items were assigned to the 11 study and test lists (i.e. the 9 critical cycles and 2 practice cycles) for each participant by randomly sampling 11 subsets of 210 items without replacement, and randomly assigning 135 items from each subset to the target category for a given cycle, and the remaining 75 to the lure category for that cycle. For the ratings condition, the 4 study and test lists (3 critical cycles and 1 practice cycle) for each participant were constructed by randomly sampling 4 subsets of 210 items without replacement, and assigning items to the target and lure categories in the same manner as in the binary condition.

Procedure

Participants were tested individually and were seated approximately two feet in front of a computer monitor. In the first session of the binary condition, participants were informed that they would be shown a list of words to study, following which their memory for the words would be tested. Participants then began the first study list, which consisted of 135 items presented randomly, one at a time, centered on the computer screen for 1 second each. This was followed by a test list containing a random selection of 75 targets and 75 lures, also presented randomly, one at a time, and centered on the computer screen. For each item, participants made an old/new response via key press.

The instructions for the initial test were randomly selected without replacement from one of the three bias clusters in combination with a specific criterion condition. The latter was sampled without replacement from one of the orders in the set {(LMS), (SLM), (MSL)}. In other words, the instructions for the first test corresponded to a condition sampled randomly from the full set of 9 bias conditions. In the instructions for a given condition, participants were asked to keep either their hit rate, false alarm rate, or overall 'Old' rate within bounds which varied according to the condition. The specific bounds stated for each condition are listed in Table 1. Following the instructions, participants advanced through the 150-item test, during which they were given feedback at several unpredictable intervals. The feedback consisted of the message "Too Strict," "Good," or "Too Lax," depending on the condition and response rates produced within the immediately preceding interval.

At the end of the test list, participants were informed that they could take a short break if desired, and that they should advance to the next study-test cycle when ready, via

key-press. The procedure for the following study-test cycle was the same as for the previous cycle, with two exceptions. First, participants were told that none of the words from any of the previous lists or cycles would be presented during the current cycle. Second, the test instructions corresponded to the next criterion condition from the order that was sampled for the cluster used in the session. The procedure for the third cycle was analogous to the procedure of the second cycle. Participants were expected to return to the lab for a second and third binary session, each separated from the preceding session by 1 to 2 days. For the second and third binary sessions, the experiment proceeded as in the first binary session, but participants were assigned to a new condition sampled without replacement from the set of remaining clusters and criterion orders.

The procedure for the ratings condition was the same as for the binary condition, but consisted of only one session and did not include any instructions relating to the response rates or feedback. Participants were instead asked to make a confidence rating on a scale of 1 to 3 where a 1 corresponded to 'Not at all confident,' a 2 corresponded to 'Moderately confident,' and a 3 corresponded to 'Very confident.' As the ratings 1-3 were given following both 'old' and 'new' responses, they were subsequently recoded as a 6-point scale ranging from a high confidence 'old' to a high confidence 'new' response in the analyses.

To help participants orient to the task, the experiment began with a separate session during which they were given practice using the ratings and binary methods. This session consisted of three study-test cycles. Two cycles used the binary procedure with instructions randomly sampled from the set of 9 in Table 1, the only constraint being

that the same condition was not repeated within the practice session. The remaining cycle used the ratings procedure.

Results

The data from the critical binary and ratings conditions were used to construct ROC curves for all individual participants and group ROCs for each condition. The resulting functions are plotted in Figure 4. The 2HT and unequal-variance SDT models were fit to the ROCs by minimizing the maximum likelihood estimator G^2. An additional model called MPTC, which was proposed by Klauer and Kellen (2010; 2011), was also fit to the ratings data using the same method. The flexibility of this model was previously evaluated by Dube, Rotello, and Heit (2011), and a similar MPTC analysis was conducted by Dube and Rotello (under review). As shown in Figure 3, the MPTC model is the same as the 2HTM but extends the detect states to the high, medium, and low confidence levels for a given item type using the parameters s_l, s_m, and s_h. As these parameters must sum to 1, the MPTC model has a total of 9 free parameters, two more than the 2HT and SDT models for ratings data. The equations for all three models can be found in Appendix A.

The values of the fit statistics G^2, *AIC*, and *BIC* are reported in Tables 2-4, and the best-fitting parameter values for each model fit to the group data are reported in Table 5. *AIC* and *BIC* are fit statistics that are useful for comparing models with different numbers of free parameters (Akaike, 1973; Schwartz, 1978). Both statistics contain a term that measures goodness of fit, and a term that constitutes a penalty for each free parameter. The statistics differ only in that the penalty for model complexity is more severe in *BIC*.

In all three statistics, smaller values indicate a better fit for a given model. As the SDT and 2HT models have an equal number of parameters, the fit of these models to each ROC curve can be compared using G^2.

Goodness of Fit: Binary Data

Considering first the binary ROCs, the data in Figure 4 appear to be quite variable, and do not obviously conform to either a line or a curve. In several cases (e.g. Participants 2, 8, 10, and 14), fairly large changes in accuracy are apparent. For instance, Participant 8 produced operating points differing in P_r by as much as .43, and in d' by as much as 1.16. This indicates a failure to meet the central assumption of ROC analysis that the operating points differ in response bias only. Related to this, the data in Table 2 show that both models failed to describe nearly half of the individual participants' data (including the participants mentioned above), and several fairly large departures in G^2 are apparent. Of the 14 participants in Table 2, one case was a tie (Participant 4), and 8/13 remaining cases were better fit by SDT than the 2HTM. The magnitude of the fit difference ($G^2_{2HT} - G^2_{SDT}$) is fairly small in most cases (mean difference = .66), though a one-sample *t* test of the difference favoring SDT approached significance: $t(13) = 2.13$, $p = .053$. Summing the G^2 values for the two models, the results show a slightly better fit for SDT (210.74 vs. 219.97), although here both models depart from the data even when the compromise-power criterion adopted previously by Bröder and Schütz (2009) is applied ($\chi^{2*}(98) = 170.39$). Surprisingly, the group ROC is slightly better fit by the 2HT than SDT, which conflicts with the individual-participant data. Though this may indicate averaging over participants has distorted the data, it is important to note that the

magnitude of the difference ($G^2_{2HT} - G^2_{SDT} = -1.15$) is consistent with the generally small differences in model fit obtained at the individual-participant level. Additionally, it is clear in the Group plot that the data are clustered closely together, covering a range that even in the ratings data does not look particularly useful in discriminating between the models. This is consistent with the model recovery simulations reported by Dube et al. (2011) that showed these models are difficult to distinguish when bias does not vary widely from point to point.

Together, the individual and group results show the models are performing about equally well in describing the binary ROCs, and frequently they both depart from the data. In considering these results along with the failure to maintain constant accuracy in the individual plots, a likely conclusion is that the differences in fit are random and do not represent true individual differences in ROC form.

Goodness of Fit: Ratings Data

The ratings data are plotted in Figure 4, and the corresponding fit statistics are reported in Table 3. Some participants did not use every rating category, which resulted in fewer freely-varying response rates (9 rates for each) than there are parameters in MPTC. Thus, the summed fits are based only on those 10 participants who used the full scale. The remaining participants were fit with the SDT and 2HTM using the same parameters as for the other participants, but with only 2 *df* contributing to the G^2 tests.

It is clear from the plots in Figure 4 that the ratings functions are curved. This result is not surprising, given the substantial evidence for curved ratings ROCs in the literature. Correspondingly, for 13 out of 14 cases in Table 3 the SDT model

outperforms the 2HTM, which predicts linear ratings ROCs in this case. This result is significant at the .01 level according to a sign test. The summed G^2 is 69.57 for SDT and 303.58 for the 2HTM if all 14 cases are considered, and the group data are likewise better fit by SDT (9.43 vs. 200.68).

More of interest is the fit of the threshold model when it is augmented to handle curvature. Though this model, MPTC, provides the best fit in terms of G^2, the picture reverses somewhat when the greater complexity of this model is factored in. Here, the results show that SDT provides the best description of the data in 7 out of 10 cases in *AIC*, and in 9 out of 10 cases in *BIC*. The latter result is significant at the .05 level according to a sign test. The somewhat harsher penalty for complexity in *BIC* produces results that are more strongly in favor of SDT, as is evident in a comparison of the results in summed *AIC* (13679.47 vs. 13698.39) and summed *BIC* (13967.12 vs. 14068.21). Similar results are also obtained at the group level: both *AIC* (20818.71) and *BIC* (20865.94) are lowest for SDT. These results indicate that even when the response stage of the threshold model is augmented to produce curvature, the results still do not map onto the observed functions as well as those predicted by SDT. This suggests that curvature in ratings ROCs is not due to complexity in the response stage, but may be a result of the processes underlying recognition judgments.

Goodness of Fit: Binary and Ratings Data

To more closely examine the correspondence between the ratings and binary ROCs, both datasets were fit simultaneously by SDT and the 2HTM (at the group and individual-participant levels). Specifically, an SDT model with 16 parameters (one slope

parameter, one mean, and 14 criteria) was compared to a threshold model that combines the predictions for the binary and ratings functions. For the latter model, a single set of detection parameters was used, but the parameterization of the response stage differed depending on the ROC type. For the binary data, the 2HTM response stage was adopted, but for the ratings data the response stage of MPTC was used. This threshold model has a total of 18 parameters (two detection parameters, 2 freely varying *s* parameters, and 14 freely varying *b* parameters), two more than the SDT model applied to the same data.

Model-predicted ROCs for the full dataset are shown in Figure 4 as solid lines for the binary data (both models) and ratings data (SDT). The 2HTM predictions for the ratings data are shown as crosses. The observed data in Figure 4 do not show a strong correspondence between the ratings and binary ROCs. This may be due in part to the fact that the binary data are not cumulated as are the ratings data, and thus may vary more widely than is possible in the ratings data (i.e. *H* and *F* must increase monotonically in the ratings data, but not in the binary data). Unfortunately, there are also three participants (3, 4, and 9) for whom the ratings functions fall higher in the space than the corresponding binary data, indicating higher accuracy in the ratings session. Though this may be due to random fluctuations across sessions, it also creates a structural disadvantage in SDT for these participants, as only a single function is implied by this model. More troublesome is the fact that both models are rejected in 12/14 datasets in Table 4. Though absolute fits are less informative in this analysis than comparative fits, the number of observations driving G^2 is less likely to produce deviations in the individual fits where these rejections are occurring, than in the group and summed fits. This indicates a lack of correspondence between the binary and ratings data that is not

due to the sort of difference assumed by MPTC, in which case the extended 2HTM would have provided an adequate fit to the complete ROCs for most participants. The high rejection rates observed for both models indicate these data either do not provide a valid way of testing their assumptions, or that the assumptions of both models are incorrect. Given the fluctuations in accuracy apparent in the plots, the most conservative conclusion is the former.

Perhaps not surprisingly given the individual fits, the results in Table 4 do not conclusively favor either model. While both models provided a good fit to the group data according to the compromise criterion ($\chi^{2*}(12) = 84.67$ for SDT, $\chi^{2*}(10) = 81.47$ for the 2HTM), summed G^2 exceeds the corresponding compromise criterion in each case ($\chi^{2*}(164) = 263.54$ for SDT, $\chi^{2*}(136) = 233.28$ for the 2HTM). Equally troublesome is the fact that the group results show a better fit for SDT in G^2 though summed G^2 is actually lower for the 2HTM (25.71 vs. 26.69). As for the binary data, the difference in fit now favoring the 2HTM is small for most participants (mean difference = -3.64) though a test of the difference falls just short of significance: $t(13) = 2.13, p = .053$. The binary data make up more than half of the data being fit in each plot, however, and a reversal was previously observed in the fits to the same binary data. This suggests the pattern is spurious in nature, rather than an effect of averaging per se.

Considering *AIC*, MPTC outperforms SDT in 7 out of 14 cases, though the sum is actually lower for SDT: $\Delta AIC = 4.99$. In terms of *BIC*, which includes a larger penalty for complexity, SDT performs better in 13 out of 14 cases. The latter result is significant at the .01 level according to a sign test. Summed *BIC* is also lower for SDT than MPTC (43022.55 vs. 43181.44). Similar results were obtained at the group level: 44717.64 for

SDT vs. 44738.89 for MPTC. The differences in *BIC* for the summed and group comparisons are 158.89 and 21.25, respectively, and constitute strong support for SDT whether simple rules of thumb or evidence ratios are considered (Burnham & Anderson, 1998; Wagenmakers & Farrell, 2004).

To summarize, while the results for the ratings data favor SDT, the binary data are not diagnostic. If the difference between the two classes of ROCs was due to artifactual curvature introduced via the ratings procedure, however, the extended 2HTM should have provided a better description of the full dataset than SDT. Unfortunately, the results from the latter analysis do not discriminate between the models, and in almost every case both models failed to adequately describe the individual-participant data. Therefore, no strong conclusions can be drawn on the basis of the goodness-of-fit results. To more closely examine the differences between the models, it may be helpful to consider the parameter values obtained in these fits.

Parameter Values

As shown in Table 5, the SDT parameters μ_o and σ_o are more or less stable across the conditions, though μ_o increases somewhat in the ratings condition relative to the binary condition. This results in a slightly higher sensitivity estimate for the ratings data: d_a = .93 vs. .83 for the binary data. Changes of a similar magnitude are observed for the two sensitivity parameters of the 2HTM: p_o increases by .09 and p_n decreases by .11 from the binary to the curved ratings fits. Although these numbers are similar in magnitude to the change in d_a for SDT, it is important to keep in mind the difference in scale for these sensitivity estimates. For instance, a difference of .10 in P_r can correspond to a

difference in d' that is greater than 1.00, depending on overall bias. Thus, these changes may be taken as a failure of the assumption in MPTC that the detect states are unaffected by response mappings (Klauer & Kellen, 2010; 2011), but the fact that changes of some magnitude were apparent in both models and in the same direction indicates accuracy may in fact be slightly higher in the ratings session. It is difficult to draw any strong conclusions here, given the general lack of correspondence between the ratings and binary data and the poor fits for both models.

Discussion

The results from Experiment 1 failed to show a clear correspondence between ratings and binary ROCs, although the lack of correspondence in this dataset was not due to systematic differences in ROC form as hypothesized by Bröder and Schütz (2009). Here, the difference was essentially that the ratings data exhibited a systematic form while the binary data showed wide between- and within-participant variability and did not conform to either SDT or the 2HTM. As is clear in Figure 4, several participants show near-chance accuracy in a subset of the binary sessions. It is also clear in the group plot and the parameter values for both models that the ratings data show slightly higher accuracy than the binary data, and that for some individuals there are differences in accuracy for the ratings and binary functions (e.g. Participants 3, 4, and 9). This may indicate a buildup in proactive interference or fatigue across the sessions in this experiment. Across session interference effects of this sort have been widely documented in recall tasks, though they are typically less marked in recognition (Underwood, 1957; McGeoch & Irion, 1952). Accuracy overall is not spectacular either,

which could indicate within-session fatigue or interference in place of or in addition to effects across sessions.

It is thus desirable to know whether accuracy changes over the course of the experiment. These data (Hits, False Alarms, d_a, and P_r) are plotted as a function of time of test, in Figure 5. There do not appear to be any consistent effects of time of test on accuracy either within- or across-tests in this experiment, although a small effect across sessions and a possible inter-session effect on sessions 1 and 2 appear for d_a. Another way to assess fatigue effects is to consider the data in Table 2 separately for participants who completed the ratings task during either the first or final session. Participants who performed the ratings task last are marked with a cross. As can be seen in the table, 3/7 of these participants were poorly fit by both models, which is the same proportion that were poorly fit by both models when all 14 cases were considered. Several of these participants also show low accuracy in one or more binary operating points (Participants 7, 10, 12, and 13). This suggests the variability in the present dataset was not due to fatigue or interference effects.

A final possibility is that the current results are a function of the task itself. As noted above, the criterion manipulation used here was modeled on a previous design by Egan et al. (1959) for a tone detection task. Yet it is well known that participants in recognition experiments often show a reluctance to shift their criteria in response to experimental manipulations when base rates and incentives are constant (Verde & Rotello, 2007; Stretch & Wixted, 1998). Though more recent work suggests that participants will shift criteria if the manipulation occurs between lists rather than within-lists (Hirshman, 1995; Hockley & Niewadomkski, 2001; 2007; Verde & Rotello, 2007),

this fact does not ensure that between-list criterion shifts are accomplished in a uniform fashion between- or within-participants, or that it is an easy task to accomplish in general. In the present experiment, an additional difficulty presents itself: an equal number of targets and lures was presented to each participant on each test list, despite instructions to maintain response rates that sometimes require extreme biases. This allows for the possibility of a conflict between metacognitive assessments of memory strength and the desire to conform to instructions regarding the response rates. In Experiment 2, binary ROCs are constructed using an actual base rate manipulation, as in Bröder and Schütz (2009). If the results in Experiment 1 were due to the task itself, the binary data in Experiment 2 should show greater stability and hopefully allow a more definitive test of the assumptions of SDT and the 2HTM.

CHAPTER III

EXPERIMENT 2

The goal of Experiment 2 was to provide a replication and closer examination of the procedure used by Bröder and Schütz (2009). There were two conditions in the present experiment. In one condition, the proportion of old items at test was varied across 9 levels, within-participants, similar to the manipulation used by Bröder and Schütz (a within-participants version with 5 levels was used in their Experiment 3). Participants were informed of the base rate manipulation, and were asked to try to match the base rates at test. Participants were also given feedback comparing their actual 'old' response rate to the desired rate at each of several pre-determined intervals during a given test. In the other condition, participants were additionally asked to follow each old/new response with a confidence rating. These manipulations allowed a direct comparison of ratings and binary ROCs in the present experiment, as well as a comparison with the results of Experiment 1.

Method

Participants

Thirty-eight psychology students at the University of Massachusetts participated; they received either course credit or payment in exchange for their participation. Several datasets were lost or excluded due to experimenter error (1), programming error (1), failure to complete all of the sessions (3), failure to use more than one confidence rating (1), or chance responding (3). Of the remaining 29 participants, fifteen were assigned to the binary condition and 14 were assigned to the binary plus ratings condition.

Design

The binary condition of Experiment 2 used a 2 (old or new test item) x 9 (criterion condition: strict, medium, or lax, in 3 bias clusters) repeated-measures design. Thus, the design is the same as in Experiment 1, except that the criterion conditions in this experiment were defined by the proportion of items on the test that were actually old, which varied across 9 levels as illustrated in Table 6.

The binary plus ratings condition used the same design as the binary condition, but included instructions to the participants to follow each old/new decision with a confidence rating on a 3-point scale ranging from 'not at all confident' to 'very confident.'

Stimuli

The same stimulus pool was used as in Experiment 1. For the binary condition, items for the to-be-studied and corresponding target categories were assigned to the 9 critical study and test lists for each participant by randomly sampling 9 subsets of 135 items without replacement for the study lists, and randomly assigning some (or all) of the items from a given subset to the target category for the corresponding test list. The precise number of studied items presented on a given test is listed in Table 6, by condition. For each of the 9 test lists, participants saw a total of 150 items. Hence, the number of lures for each list was complementary to the number of targets, as shown in Table 6. The lures for each list were likewise randomly selected without replacement

from the stimulus pool. Stimuli were also assigned to the three study-test cycles of an initial practice session, as in Experiment 1. In the present experiment, the three cycles used three randomly selected conditions from Table 6, the only constraint being that no condition was repeated within the practice session.

Procedure

The procedure for the critical sessions was the same as in Experiment 1, with the exceptions that 1) the precise number of targets and lures at test varied across conditions and participants were informed of this, 2) the criterion instructions and feedback were replaced with instructions and feedback that stressed the base rate manipulation, and 3) the ratings condition of Experiment 1 was replaced with the binary plus ratings condition, and used a different sample of participants than the present binary condition. The binary plus ratings condition only differed from the binary condition in that it required participants to follow each old/new decision with a confidence rating on a scale of 1-3. In the first session of Experiment 2, all participants were given practice on the procedure appropriate to their condition (binary or binary plus ratings). The practice session consisted of 3 study-test cycles using three randomly sampled conditions from Table 6, the only constraint being that no condition was repeated within the practice session.

Results

The critical binary and ratings data from the binary and binary plus ratings conditions were used to construct ROC curves for all individual participants and group ROCs for each of the three datasets. The 2HT and unequal-variance SDT models were fit

to the binary ROCs from both conditions and the ratings ROCs from the binary plus ratings condition (collapsing across the binary conditions)[3]. In addition, the MPTC model was fit to the ratings ROCs from the binary plus ratings condition, and was compared to SDT in simultaneous fits of the binary and ratings data. As in Experiment 1, both the group and individual-participant ROCs were fit with all three models. The values of the fit statistics G^2, *AIC*, and *BIC* are reported in Tables 2 and 7-8, and the best-fitting parameter values for each model fit to the group data are reported in Table 9.

Goodness of Fit: Binary Data

Considering first the binary ROCs, the data in Figure 6 (filled circles) are quite variable, and do not clearly imply either a line or a curve. Large changes in accuracy are also apparent in several of the plots. For instance, Participant 6 produced operating points differing by as much as .31 in P_r, and in d' by .83. These changes in accuracy indicate the ROCs obtained in this experiment may not be valid, as ROCs assume that only bias varies across the points. Unlike Experiment 1, however, the data in Table 2 indicate that most participants are well-described by at least one model, and the departures in G^2 that are apparent are relatively small in magnitude for half of these cases (all *p*-values > .01). Interestingly, the departures tend to be larger in magnitude in the binary plus ratings condition. As only 4 out of 14 participants in this condition show large departures in fit, however, it is not unlikely that the difference across conditions is entirely spurious. If the binary data from both conditions are considered together, only

[3] The group ratings ROCs were unweighted averages of the participant x condition ratings ROCs. Weighted averages in this analysis may yield group ROCs that are above chance even if the corresponding participant x condition ROCs are at chance, which would necessarily inflate G^2 in the simultaneous fits and would not be representative of the functions they collapse across.

8/29 = 28% show a departure of any sort for at least one model, as opposed to the 50% rate in Experiment 1. As is apparent in Table 2, there is also greater agreement and better fits overall when the individual, summed, and group G^2 values are considered.

Looking first at the individual data, Table 2 shows that 9 out of 13 participants in the binary condition were better fit by SDT than the 2HTM (the models were tied for participants 7 and 10). For the binary data from the binary plus ratings condition, 9/14 participants were better fit by SDT. Although the magnitude of these fit differences in neither case approaches significance, the more adequate sample size afforded by the full binary dataset shows an advantage for SDT, $t(28) = 2.07, p<.05$. Summed G^2 also shows a better fit for SDT in both the binary (135.15 vs. 141.36) and binary plus ratings conditions (176.72 vs. 187.16). While summed G^2 indicates a departure for both models, only the 2HTM shows a departure when the compromise-power criterion ($\chi^{2*}(105) =182.62$) is applied. The group data for these two binary conditions are consistent with the individual-participant data: G^2 is in each case smaller for SDT than the 2HTM (3.83 vs. 7.38, for Binary; 15.59 vs. 16.99 for binary plus ratings). Although the group fits for the binary plus ratings condition indicate a departure for both models, here neither model exceeds the compromise criterion ($\chi^{2*}(7) =63.48$). Together, these results show that SDT provides a better account of the binary data than does the 2HTM.

Goodness of Fit: Ratings Data

The ratings data are plotted in Figure 7 and the corresponding fit statistics are reported in Table 7. Several participants failed to use the full ratings scale, producing just enough data to fit SDT and the 2HTM for 13 participants without saturating those

models. Participant 8 produced fewer freely-varying response rates than there are parameters in any model, and was excluded from this analysis. The remaining 8 participants used the full ratings scale and were fit by all three models.

The ROCs in Figure 7 appear curved for most participants, the possible exceptions being Participants 1 and 6. Unexpectedly, several participants were not well-described by SDT (6/13). Several ROCs show a close clustering of the operating points, producing as few as three useful points (the ROCs for Participants 1, 2, 4, 9, 12, and 14 all show this pattern). However, only two of these participants (Participants 4 and 12) produced divergent fits in SDT, and in any event the spread and number of functional operating points is generally an issue for model selection rather than absolute fit (Dube, Rotello, & Heit, 2011). Perhaps a more likely reason for the misspecifications in SDT is that these ROCs were estimated using unweighted averages. Though this method was chosen in order to avoid distortions from averaging over conditions differing in old item base rates, the results may still not be representative of the participant x condition ratings ROCs contributing to the averages. I will return to this issue in the *Discussion* section below.

Despite the issues in overall goodness of fit, the present results are typical in that 12 out of 13 cases were better fit by the SDT model than the linear 2HTM. Summed G^2 is 166.74 for SDT and 513.16 for the 2HTM when all 13 participants are considered (both values exceed the compromise criterion $\chi^{2*}(33) = 91.85$). The group data are also better fit by SDT (46.93 vs. 399.84). Here, only the 2HTM fit exceeds the compromise criterion for the group data ($\chi^{2*}(3) = 49.20$). When all three models are considered, MPTC provides the best fit in 6 out of 8 cases, in summed G^2 (53.12; significant

according to the compromise criterion with 8 *df*), and in the group fit (G^2 = .38). When the greater complexity of MPTC is considered in the fits, this model still provides the best description of the data. This results holds at the individual-participant level (6/8 cases for both *AIC* and *BIC*), in summed *AIC* and *BIC* (30597.82 and 30899.88, respectively), and in the group fits (*AIC* = 63954.57, *BIC* = 64025.19). Again, these results should be interpreted cautiously, as the ratings data may have been distorted by the use of unweighted averages.

Goodness of Fit: Binary and Ratings Data

To more closely examine the correspondence between the ratings and binary ROCs, both datasets were fit simultaneously by SDT and the 2HTM (at the group and individual-participant levels). An SDT model with 16 parameters was compared to a threshold model with 18 parameters to combine the predictions for the binary and ratings functions. Model-predicted ROCs for the full dataset are shown in Figure 7 as solid lines for the binary data (both models) and ratings data (SDT). The 2HTM predictions for the ratings data are shown as crosses. The observed data in Figure 7 do not show a strong correspondence between the ratings data and the binary data. More troublesome is the fact that both models are rejected on nearly half of the datasets (6/14 cases in Table 8). The finding of poorer fits for both models in the simultaneous fits suggests a discrepancy between the binary and ratings data that is not due to the response stage alone (as assumed by MPTC). This may be a further consequence of the unweighted averages used in constructing the ratings functions.

The results in Table 8 do not conclusively favor either model. The 2HTM outperforms SDT in *AIC* in 8 out of 14 cases, and the sum is also lower (68724.37 vs. 68786.42). In terms of *BIC*, which includes a larger penalty for complexity, SDT performs better in 11 out of 14 cases. The latter result is marginally significant according to a sign test ($p = .057$). Summed *BIC* is also lower for SDT than the 2HTM (70102.3 vs. 70205.4). Both models failed to provide an adequate account of the group data, $G^2(16) = 82.80$, $p < .001$ (SDT) and $G^2(18) = 46.00$, $p < .001$ (MPT), although both fits are adequate by the standards of the compromise-power criteria ($\chi^{2*}(12) = 84.67$ (SDT); $\chi^{2*}(10) = 81.47$ (2HTM)). Consistent with the pattern observed in the individual-participant analysis, the 2HTM performed better than SDT in group *AIC* (85842.44 vs. 85875.25), but contrary to the summed *BIC* results, the 2HTM also outperformed SDT in the group *BIC* (85996.16 vs. 86011.89). This may suggest distortion from averaging over participants, though as noted previously the ratings data in the present experiment were poorly fit, and this may be due to the use of unweighted averages.

Parameter Values

As can be seen in Table 9, the SDT parameters μ_o and σ_o do not vary greatly across the conditions, though μ_o decreases somewhat in the simultaneous condition relative to the binary condition (from binary plus ratings). This results in a slightly lower sensitivity estimate for the full dataset: $d_a = .69$ vs. .77 for the binary data. A similar result holds in the comparison of the same binary data to the ratings data from that condition, which is consistent with an effect of the averaging method for ratings ROCs. Changes of a larger magnitude are observed for the two sensitivity parameters of the

2HTM: p_o increases by .11 and p_n decreases by .24 from the binary to the complete fits. Similar results hold for a comparison of the same binary data to the curved ratings parameters. It is important to keep in mind the difference in scale for these accuracy parameters, as fairly small differences in P_r may translate into d' differences as large as 10 times the difference in P_r. This variability across response formats may indicate a failure of the assumption in MPTC that the processing stage is unaffected by response mappings. All the same, it is difficult to draw any strong conclusions about parameter invariance or true changes in accuracy. In this experiment, there was a general lack of correspondence between the ratings and binary data that may stem from the relatively poor fits to the ratings data.

Ratings ROC Slopes

A final consideration is the form of the ROCs from which the binary operating points were sampled. Previously, researchers in perception (Schulman & Greenberg, 1970; Treisman & Faulkner, 1984) and recognition (Van Zandt, 2000) have demonstrated that the slope of the ratings ROC tends to increase as the proportion of old items increases at test. If this is in fact a general property of ROC functions, and not an epiphenomenon of ratings-based approximations to those functions, then one might expect binary ROCs to deviate from a curve depending on what part of the underlying function a given point is sampled from. Unfortunately, clear conclusions cannot be drawn in the absence of ratings data, as a fairly narrow range of binary operating points is available to construct the functions. For instance, three Strict points from the .10, .20, and .30 conditions would provide one of the three slope estimates in this analysis.

To examine possible slope changes, the ratings ROCs from the .20, .50, and .80 base rate conditions were fit with the SDT model. These fits produced slope estimates of .69, .80, and .77 for the .20-.80 conditions. In other words, the present experiment did not produce slope changes of the magnitude reported previously (e.g. Van Zandt, 2000, used a similar design and reported an average increase of about .20 as P(Old) increased from .20 to .80). Not surprisingly, restricting the SDT model such that all three ROCs were fit with a single mean and variance resulted in a fit not significantly worse than that of the full model: $\Delta G^2(4) = 4.58, p = .33$.

The slope analysis was also conducted by fitting the SDT model to the individual-participant ratings ROCs from the three conditions and comparing the slope parameters in a repeated measures ANOVA. In this analysis, three participants were excluded as they failed to spread their ratings out sufficiently for the model to converge. For the remaining 11 participants, the averages obtained were .67, .73, and .81 for the .20-.80 conditions. These differences were not significant, $F(2,20) = 1.56, p = .23$. Although power was quite low (.292), so was effect size (Cohen, 1988): $\eta^2_p = .135$. It is possible, given the small sample and wide sampling variability of zROC slopes (Macmillan, Rotello, & Miller, 2004) that this pattern is not representative of the population in either direction or magnitude. Additionally, the analysis rests on the assumption that these ratings ROCs provide a good approximation to the underlying functions, which may be questioned but is consistent with previous conclusions regarding ratings and binary ROCs.

Discussion and Comparison with Experiment 1

To summarize, the binary and simultaneous fits were somewhat better for both models in the present experiment than in Experiment 1, which may suggest an effect of the specific task in the former experiment. In Experiment 2, the binary data were fit better overall by SDT than the 2HTM. This is consistent with the trend observed for binary data in Experiment 1, though greater consistency across the levels of analysis was observed in Experiment 2. Nonetheless, the ratings and simultaneous fits were still far from perfect. In Experiment 2, the fits to the ratings data were surprisingly poor for both models. It could be that the unweighted averaging used to construct ratings ROCs in the present experiment has distorted the results.

To explore this possibility, the ratings data from the .50 condition were considered. This is consistent with Experiment 1 (and typical ratings designs) where extreme biases were not introduced and equal numbers of old and new items were used. The results are shown in Table 10. As every participant in the ratings-only analysis produced fewer than 10 freely-varying response rates, only SDT was fit to the ratings data. Of these, 5 were excluded from the ratings-only fits as they did not produce a sufficient number of freely-varying response rates for SDT to be fit without saturating the model. Three of these five produced fewer than 3 distinct operating points in the ratings data and were thus excluded from all of the analyses in Table 10. The binary and ratings data (again from the .50 condition) were also fit as one dataset. Here, both SDT and the 2HTM were fit to the data, as in Experiments 1 and 2.

A number of changes are apparent, one being the fact that while several ratings ROCs were misspecified by SDT previously (6/13 vs. only 3/14 in Experiment 1), now

only 2/9 are misspecified, which is more in line with Experiment 1. Considering the fits to the binary and ratings data, it can be seen that only 2/11 are poorly fit by both models, in contrast to the analysis using unweighted averages where 6/14 cases were not well described by either model. Both models were rejected in summed G^2 however, even according to the compromise-power criteria of 225.08 (SDT) and 200.09 (2HTM). The group fits, which also departed from the data according to the standard analysis, were adequate according to the compromise criteria of 61.26 (SDT) and 58.31 (2HTM). While previously the 2HTM showed a slight advantage over SDT in *AIC* and *BIC*, the picture again reverses somewhat when the unbiased ratings data are used. Here, 6/11 cases are better fit by SDT according to *AIC*, and the differences favoring SDT tend to be of a larger magnitude than those favoring the 2HTM. This results in a lower summed *AIC*, and the group fit is also lower for SDT. In *BIC* 9/11 participants are better fit by SDT than the 2HTM, which is marginally significant according to a sign test ($p = .065$). Though the latter result is weaker in the present analysis, the overall pattern in *AIC* and *BIC* is now consistent with that of Experiment 1.

Taken together, these results suggest the advantage observed previously for the 2HTM in Experiment 2 was due to a distortion from averaging across ratings ROCs from different base rate conditions. When the unbiased ratings data are combined with the binary data in Table 2, the results show that SDT provides the best description of the binary and full datasets in Experiment 2. This suggests that some of the variability in Experiment 1 was due to the task, which required criterion shifting in response to instructions despite full knowledge that the base rates were constant. This may have induced a conflict between participants' metacognitive assessments of probe strength and

the feedback delivered during the course of the test. Another possibility is that trial-by-trial criterion shifting is more 'natural' to participants when the base rates are manipulated (or assumed to be manipulated), rather than experimenter-defined variables such as 'strength' (Stretch & Wixted, 1998). These possibilities will require further attention than can be given at present, however, and must be considered speculative.

CHAPTER IV

GENERAL DISCUSSION

The results from Experiments 1 and 2 do not support the idea that binary ROCs are linear. To the contrary, in the only outcome where the results were decisive one way or the other, the binary data were consistent with SDT and not the 2HTM. In Experiment 1, the results did not clearly favor either model, but nearly half of the individual cases were poorly fit by both models and the binary ROCs showed large fluctuations in accuracy. This indicates that there may have been issues specific to the task, which was adapted from an earlier design for tone detection (Egan et al., 1959). In Experiment 2, the base rate manipulation favored by Bröder and Schütz (2009) was used, and the results were usually well-described by at least one model. In that experiment, the binary data were better fit by SDT. Further, both experiments demonstrated that when ratings data are included in the fits the results are slightly better described by SDT than MPTC, even though the latter model can produce curvilinear ratings ROCs. If the curvature in the ratings data was due to variability in how the ratings were used by participants, the extended 2HTM should have consistently provided a better description of the full dataset than SDT. This was not so. In fact, the SDT model, which predicts only one ROC function for the binary and ratings data, performed somewhat better than the extended 2HTM in both experiments. This suggests the curvature that is typically observed in ratings ROCs is a general property of ROCs, consistent with the assumptions of SDT.

These findings, though they may not convince all readers, are nonetheless consistent with a growing body of literature including the recognition data in Bröder and

Schütz's meta-analysis, the implied base rate data of Dube and Rotello (under review), the binary ROCs collected by Dube, Rotello, & Heit (2010; 2011), and the many studies conducted by researchers in perception (see Dube & Rotello (under review), for review). These studies are opposed in their conclusions by only one new study by Bröder and Schütz that did not clearly favor either model. Thus, the bulk of the research on the topic to date converges on one conclusion: binary ROCs are curved and similar in form to ratings ROCs.

The curvature documented for binary and ratings ROCs has important implications for how recognition accuracy is assessed. As Kinchla (1994) has demonstrated, two conditions may be found to differ in accuracy when in fact they differ only in bias if threshold statistics are used when the underlying ROCs are curved. At first glance, this may seem like a scenario that would not survive replication. However, Rotello, Masson, and Verde (2008) have shown via simulations that Type I errors in this scenario are persistent and even become more likely with larger samples. This implies that such errors could potentially survive replication, high-power designs, and extensive theoretical scrutiny.

Empirically, Dube et al. (2010, 2011) have documented precisely this state of affairs in the reasoning literature. Their study examined the belief bias effect (Evans, Barston, & Pollard, 1983; Klauer, Musch, & Naumer, 2000), the finding that accuracy in evaluating deductive arguments is greater when those arguments imply unbelievable conclusions (e.g. 'All pets are dogs') than when they imply believable ones (e.g. 'Some pets are dogs'). This effect has been a source of theoretical debate in the reasoning literature for nearly 30 years. Importantly, accuracy in the belief bias task is typically

measured with a contrast of P("Valid"|Valid) and P("Valid"|Invalid), in other words *H* – *F*, computed separately for believable and unbelievable arguments that vary in validity status. Dube et al. examined ROC curves for the belief bias task and found them to be curvilinear, consistent with previous results in perception and recognition. More importantly, the operating points for believable and unbelievable problems fell on a single function, indicating a difference in response bias only. They concluded the apparent accuracy effect that has driven nearly 30 years of theoretical work on belief bias was due to erroneous assumptions in how it was measured. Those erroneous assumptions (linearity and unit slope in ROC space) are the same ones made by the double high-threshold model with equal detection parameters, as noted previously.

Threshold models such as the one assumed by *H-F* are not only prevalent in studies of reasoning. These models, often referred to as *multinomial processing tree* (MPT) models, have also been applied in studies of social cognition (Sherman, 2006), item recognition (Bröder & Schütz, 2009), and, perhaps most frequently, source monitoring (Batchelder & Riefer, 1990; Bayen, Murnane, & Erdfelder, 1996; Bröder & Meiser, 2007; Klauer & Kellen, 2010). In source monitoring experiments, participants study items that are presented in different contexts. For instance, they may be spoken in male or female voices, presented on the top or bottom of the computer screen, in italic or normal fonts, and so on. Participants are then given a recognition test in which they must not only make an old/new judgment, but also a decision as to the source of the item (e.g. whether a test word was originally spoken in a male or female voice). These experiments have the potential to produce more data than typical item recognition experiments, which can also complicate the measurement process. For this reason, researchers began to

adopt simple MPT models as a way of separating out accuracy and bias in item and source recognition (Batchelder & Riefer, 1990; Bayen et al., 1996). Unfortunately, these models grew in popularity in advance of any serious consideration of their processing assumptions, despite an early criticism of this work by Kinchla (1994). When ROCs were eventually collected for source recognition, they were almost always found to be curved and consistent with models derived from SDT (Rotello, Macmillan, & Hautus, under review; Hautus et al., 2008; Hilford et al., 2002; Slotnick & Dodson, 2005; Slotnick, Klein, Dodson, & Shimamura, 2000; Yonelinas, 1999).

Recently, Klauer and Kellen (2010) advanced an MPT model of source recognition that can produce curvilinear ROCs, but it does so via the scale-usage parameters discussed previously. We have shown here that the basic assumptions driving both the processing and response stages of this model are inaccurate, and that the ability of Klauer and Kellen's (2010) model to mimic SDT models has been achieved through means that are not empirically justified.

This work implies that measures like $H - F$, which assume threshold models and imply linear binary ROCs, should be avoided for tasks where ROCs have been found to be curved, namely item detection (Macmillan & Creelman, 2005; Wixted, 2007), source discrimination (Rotello, Macmillan, & Hautus, under review; Hautus et al., 2008; Slotnick & Dodson, 2005), tone detection (Green & Swets, 1966), deductive reasoning (Dube et al. 2010; 2011), and inductive reasoning (Heit & Rotello, 2010; Rotello & Heit, 2009). Although many of the studies cited here only examined ratings ROCs, the agreement between the current study and previous analyses of binary ROCs validates those studies.

Though the data to date are fairly consistent, the generality of the results reported thus far is still open to question. It remains to be seen whether the results obtained with the actual and implied base rate procedures extend to other bias manipulations, an issue that the current study was designed to examine. As the results from Experiment 1 were not useful for discriminating between the models, it is still possible that results obtained with other bias manipulations will suggest different conclusions. An important goal of future research will be to explore other options such as payoff matrices or feedback manipulations. For example, misleading feedback might be used to encourage more or less liberal responding on different tests. Another option would be a refinement of the Egan et al. procedure. This of course involves questions as to why the procedure did not produce useful data in the present study.

One possibility is that the present results reflect a difficulty in shifting criteria between tests within a given session. Although previous work has found that participants are more likely to shift criteria between- than within-lists (Hirshman, 1995; Hockley & Niewadomkski, 2001; 2007; Verde & Rotello, 2007), it is still not clear that this is always a particularly easy or 'natural' thing for participants to do. In Experiment 1, between-list criterion adjustment may have been particularly difficult as it was likely to involve a conflict between accuracy maximization and responding in line with the instructions. Specifically, the participants were asked to adopt rather extreme biases though half of the items were always old, so that on some trials they may have been highly confident that an item was old despite feedback indicating they should be saying 'new' more often (or vice-versa). As previous research using binary classification tasks has shown that participants tend to focus on accuracy maximization unless given strong incentives to the

contrary (Maddox & Bohil, 2005), it is possible that they adopted strategies to minimize conflict that were not optimal with respect to the goals of the experiment.

All the same, this still would not explain why the present results disagree with those of Dube and Rotello (under review), who used an implied base rate manipulation. In that study, the instructions misled participants as to the probability of an old item at test, which was always .50 as in the present Experiment 1. The use of misleading instructions may be a quite important difference, however. Previous work in both perception and recognition has shown that participants do not depend on their own assessments of the probes when they are given misleading feedback, but that they will adjust their responses in whatever manner is indicated (Friedman et al., 1968; Han & Dobbins, 2008). Yet in the present experiment, participants had no reason to question themselves and were aware that the proportion of old items was .50. This suggests a conflict would actually be more likely in the present Experiment 1 than in the study by Dube and Rotello.

Another possibility is that the rather narrow interval specified by the instructions and feedback was problematic. In each binary condition, participants were asked to maintain response rates falling in intervals of width .10. The difficulty of maintaining a response rate in such a small interval may have led subjects to forgo the instructions altogether, producing operating points that varied randomly in their bias. This would not explain the fluctuations in accuracy that are apparent in the plots, though, or the fact that for the most part hit and false alarm rates increased with the bias. On the other hand, there is a reversal in the response rates in the group data of Experiment 1, but not Experiment 2 (see Tables 5 and 9), and there were fluctuations in accuracy apparent for

both experiments. This suggests a replication of Experiment 1 using a smaller number of conditions, each with a wider response interval, may produce more stable results.

There are other issues in the present study beyond those specific to Experiment 1, however. Chief among these is the instability apparent in the data from both experiments. As Dube and Rotello (under review) and Bröder and Schütz (2009) have previously noted, it is difficult to ensure that accuracy is not changing across the conditions in binary ROC plots. This being a central assumption of ROC analysis, it is doubly important to ensure it is met. Although Bröder and Schütz reported an ANOVA examining *d'* and *H-F* as a function of bias condition, Dube and Rotello (under review) have noted the analysis entails a circular argument. In order to know what accuracy statistics are appropriate for the data, one must examine valid ROCs. But, in order to know whether an ROC is valid, one must compare accuracy across the conditions that produced it. Dube and Rotello (under review) showed that analyses of different accuracy statistics may produce conflicting and misleading conclusions so long as their assumptions are not met (and for at least one measure this must be so).

But the key question for the present study is to what extent the fluctuations that are observed are random or represent actual changes that are tied to response strategies or some other as yet unidentified factor specific to the bias procedures that were used. Random fluctuations might be expected to produce less orderly ROCs for the binary methods than the more frequently encountered ratings methods as cumulating the ratings responses forces hits and false alarms to increase monotonically as the ratings category becomes more liberal. One possible test, then, would be to compare cumulated and binary ratings ROCs. Specifically, participants could be asked on a given test to respond

'Old' only if they are 'very sure' that an item is old, and to respond 'New' otherwise. On another test, the same participants could be asked to respond 'Old' if they are at least 'moderately sure' the item was old, and to respond 'New' otherwise. Binary ratings points could be plotted and compared with cumulated ratings and other binary ROC data to examine the extent to which accuracy fluctuations are due to procedures like the base rate and instructed criterion procedures used in Experiments 1 and 2.

Finally, it is still not clear why Bröder and Schütz's (2009) newer results do not agree with those of Experiment 2, which used the same sort of bias manipulation. One possibility, suggested previously by Dube and Rotello (under review), is that the relatively small number of either targets or lures in the extreme bias conditions produced instability in the endpoints of their ROCs. In Experiment 2, more items were used per condition, and more conditions were included, than in the Bröder and Schütz experiment. Both of these factors should have reduced any effects stemming from the endpoints. Another issue specific to Bröder and Schütz's study is the lack of individual-participant data. Without such data, one cannot be sure that their results were not influenced by distortions from averaging over participants. But in any event the explanatory burden does not rest with the current study, since Bröder and Schütz's results also conflict with their own meta-analysis results, the existing perception and reasoning data, and the data reported by Dube and Rotello (under review).

Conclusion

The present study suggests that binary ROCs are not linear, in contrast to claims made by Bröder and Schütz (2009). Although the data from one experiment were not

decisive in selecting between the SDT and 2HT models, they were most often consistent with SDT. In a second experiment, the base rate procedure used previously by Bröder and Schütz was adopted and the results were consistent with SDT and not the 2HTM. These results are consistent with a growing literature that suggests the assumptions of threshold and MPT models are violated in the areas of item recognition, source monitoring, perception, and reasoning. As previous work has shown that the use of threshold statistics such as percent correct and *H*-*F* is likely to produce persistent statistical errors, researchers should avoid these statistics or at least complement them with SDT-based measures like d' or d_a. In general, the best way to ensure a given analysis is justified is to collect ROC data. Confidence ratings should be preferred in doing so, as they provide a quick, efficient, and valid method of constructing ROCs.

Table 1. Summary of bias conditions in Experiment 1.

Exp. 1 Conditions		Manipulation		
Cluster	Criterion	Instructions	No. Targets	No. Lures
1	Strict	Keep F between .05 and .15.	75	75
	Medium	Keep "Old" rate between .35 and .45.	75	75
	Lax	Keep H between .65 and .75.	75	75
2	Strict	F: .15 to .25	75	75
	Medium	"O": .45 to .55	75	75
	Lax	H: .75 to .85.	75	75
3	Strict	F: .25 to .35	75	75
	Medium	"O": .55 to .65	75	75
	Lax	H: .85 to .95	75	75

*H: Hit rate, F: False alarm rate, "O": "Old" response rate.

Table 2. Fit statistics for binary data, Experiments 1-2.

Experiment 1			Experiment 2 Binary		Exp. 2 Binary Plus Ratings	
ID	$G^2(7)$ SDT	$G^2(7)$ 2HTM	$G2^2(7)$ SDT	$G^2(7)$ 2HTM	$G^2(7)$ SDT	$G^2(7)$ 2HTM
1	**20.79****	20.93**	**7.71**	9.10	7.94	**7.76**
2	**21.63****	23.55**	18.21*	**16.73***	**6.70**	8.97
3	**3.15**	5.73	**7.53**	7.97	**8.33**	8.75
4	8.82	8.83	10.56	**9.30**	**21.32****	21.53**
5†	**7.81**	8.86	**2.31**	4.40	**11.31**	12.44
6†	10.85	**10.32**	**15.24***	15.71*	**7.41**	8.47
7†	12.15	**11.55**	13.77	13.76	**10.43**	15.02*
8	21.89**	**21.23****	**3.03**	4.39	8.31	**7.07**
9	**8.26**	9.89	**7.35**	8.04	31.72***	**30.71*****
10†	40.30***	**39.91*****	9.19	9.20	7.52	**6.36**
11†	**15.23***	15.49*	4.54	**4.31**	**21.11****	25.00***
12†	**12.70**	15.00*	14.88*	**13.83**	**7.08**	7.45
13†	12.60	**12.45**	**8.04**	8.66	**21.74****	23.07**
14	**14.56***	16.23*	**8.78**	9.48	5.80	**4.56**
15			**4.01**	6.48		
Total	**210.74*****	219.97***	**135.15***	141.36*	**176.72*****	187.16***
Group	7.68	**6.53**	**3.83**	7.38	**15.59***	16.99*

**p<.05, **p<.01, ***p<.001.* † Denotes participants in Experiment 1 who received the ratings task on their final session. Bold values indicate the better fitting model.

Table 3. Fit statistics for ratings data, Experiment 1.

	G^2			AIC			BIC		
ID	SDT (3 df)	2HTM (3 df)	MPTC (1 df)	SDT	2HTM	MPTC	SDT	2HTM	MPTC
1	3.74	4.02	**1.29**	**1456.3**	1456.6	1457.9	**1485.1**	1485.4	1494.9
2	6.14	13.57**	**0.35**	1532.2	1539.6	**1530.4**	**1561.0**	1568.4	1567.4
3	**0.25**	14.70**	1.78	**1239.2**	1253.6	1244.7	**1267.9**	1282.4	1281.7
4	**3.07**	43.13***							
5	7.45	12.08**	**5.20***	**1352.7**	1357.3	1354.4	**1381.4**	1386.1	1391.4
6	9.54*	2.49	**2.50**	1395.7	1388.7	**1392.7**	1424.5	**1417.5**	1429.7
7	9.40**	47.09***							
8	13.43**	21.76***							
9	**0.94**	11.92**	1.90	**1298.6**	1309.6	1303.6	**1327.4**	1338.4	1340.6
10	6.12	60.05***	**1.13**	1290.8	1344.5	**1289.9**	**1319.6**	1373.5	1326.8
11	**1.13**	19.57***	2.49	**1369.9**	1388.3	1375.2	**1398.6**	1417.1	1412.2
12	0.99	11.16*	**0.25**	**1398.6**	1408.8	1401.9	**1427.4**	1437.6	1438.9
13	4.14	8.71*	**2.47**	**1345.4**	1350.0	1347.7	**1374.2**	1378.7	1384.7
14	3.23	33.33***							
Total	40.44	158.27***	**19.36***	**13679.5**	13797.1	13698.4	**13967.1**	14084.9	14068.2
Group	9.43*	200.68***	**8.65****	**20818.7**	21010.0	20821.9	**20865.9**	21057.2	20882.7

**p<.05, **p<.01, ***p<.001.* Participants 4, 7, 8, and 14 produced only 9 response rates that could not be fit with less than 9 parameters in MPTC. SDT and the 2HTM were fit to these data with the same parameters, but only 2 df. The summed values for each model include only those 10 participants who produced 10 response rates.

Table 4. Fit statistics for ratings and binary data, Experiment 1.

	G^2		AIC		BIC	
ID	SDT (12 df)	2HTM (10 df)	SDT	2HTM	SDT	2HTM
1	26.54**	**22.71***	**3156.0**	3156.2	**3243.9**	3255.1
2	34.31***	**28.62****	3242.9	**3241.2**	**3330.9**	3340.2
3	**28.37****	36.37***	**2894.0**	2906.0	**2981.9**	3004.9
4	**29.89****	30.64***	**2886.2**	2891.0	**2974.2**	2989.9
5	20.13	**14.91**	3032.4	**3031.2**	**3120.3**	3130.1
6	32.32**	**20.35***	3016.0	**3008.0**	**3103.9**	3106.9
7	28.83**	**20.97***	2781.4	**2777.6**	**2869.3**	2876.5
8	41.43***	**25.33****	3166.2	**3154.1**	3254.1	**3253.0**
9	**25.24***	26.24**	**3016.9**	3021.9	**3104.8**	3120.8
10	52.85***	**43.76*****	2904.7	**2899.6**	**2992.7**	2998.6
11	**21.66***	23.57**	**3044.2**	3050.1	**3132.1**	3149.0
12	21.12*	**16.36**	3150.7	**3149.9**	**3238.6**	3248.8
13	23.96*	**23.73****	**2821.4**	2825.1	**2909.3**	2924.0
14	**30.83****	32.94***	**2678.8**	2684.9	**2766.8**	2783.8
Total	417.48	**366.50**	**41791.6**	41796.6	**43022.6**	43181.4
Group	**25.71***	26.69**	**44587.5**	44592.5	**44717.6**	44738.9

**p<.05, **p<.01, ***p<.001.* The SDT and 2HT models were fit to Participants 4, 7, 8, and 14 with the same number of parameters as for the other cases but 11 and 9 df as these participants only produced 27 freely-varying response rates.

Table 5. Best-fitting parameter values for Experiment 1, group data.

	SDT													
Condition	μ_o	σ_o	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	d_a		
Binary	0.93	1.23	0.78	0.62	0.50	0.55	0.38	0.21	0.28	0.07	-0.01	.83		
Ratings	1.04	1.22	1.49	0.79	0.43	-0.20	-1.11					.93		
Complete (Binary)	0.96	1.22	0.80	0.64	0.52	0.57	0.40	0.23	0.30	0.09	0.003	.86		
Complete (Ratings)		1.45	0.76	0.40	-0.23	-1.13								
	2HTM													
Condition	p_o	p_n	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	s_l	s_m	s_h
Binary	0.39	0.23	0.28	0.34	0.40	0.38	0.46	0.54	0.51	0.61	0.65			
Ratings (linear)	0.31	0.09	0.07	0.23	0.15	0.25	0.25	0.05						
Ratings (curved)	0.48	0.12	0.07	0.17	0.14	0.28	0.26	0.08				0.07	0.27	0.66
Complete (Binary)	0.45	0.13	0.24	0.29	0.35	0.33	0.40	0.48	0.45	0.55	0.58			
Complete (Ratings)			0.08	0.18	0.15	0.28	0.26	0.05				0.05	0.26	0.69

* d_a is not a free parameter in SDT, but is computed using the parameters μ_o and σ_o (see Introduction).

Table 6. Summary of bias conditions in Experiment 2.

Exp. 2 Conditions		Manipulation		
Cluster	Criterion	P(Old)	No. Targets	No. Lures
1	Strict	.10	15	135
	Medium	.40	60	90
	Lax	.70	105	45
2	Strict	.20	30	120
	Medium	.50	75	75
	Lax	.80	120	30
3	Strict	.30	45	105
	Medium	.60	90	60
	Lax	.90	135	15

Table 7. Fit statistics for ratings data, Experiment 2.

	G^2			AIC			BIC		
ID	SDT (3 df)	2HTM (3 df)	MPTC (1 df)	SDT	2HTM	MPTC	SDT	2HTM	MPTC
1	6.02	10.99*	**0.14**	3341.8	3346.3	**3335.5**	3377.8	3382.8	**3371.9**
2	3.09	18.72***	**0.78**	3801.6	3817.2	**3799.3**	3838.0	3853.7	**3835.7**
3	**4.16**	105.07***							
4	20.25***	110.82***							
5	**35.13*** **	87.84***	41.17***	**4133.3**	4186.0	4139.3	**4169.7**	4222.4	4175.8
6	16.64***	18.69***	**4.19***	4230.7	4232.8	**4218.3**	4267.2	4269.2	**4254.7**
7	20.67***	10.21*	**1.86**	4522.9	4512.4	**4504.1**	4559.4	4548.9	**4540.6**
9	7.08	8.28*							
10	13.08**	39.94***	**0.50**	4106.6	4133.4	**4094.0**	4143.0	4169.9	**4130.4**
11	4.37	5.98	**0.51**	3349.7	3351.3	**3345.9**	3386.2	3387.8	**3382.3**
12	30.54***	55.09***							
13	3.04	6.18							
14	**2.67**	35.35***	3.97	**3156.3**	3189.0	3161.6	**3192.7**	3225.4	3208.5
Total	101.67***	227.72***	**53.12*** **	30642.8	30768.4	**30597.8**	30934.0	31060.1	**30899.9**
Group	46.93***	399.84***	**0.38**	63997.1	64350.0	**63954.6**	64052.1	64405.0	**64025.2**

**p<.05, **p<.01, ***p<.001.* Participants 3, 4, 9, and 12 produced only 9 response rates and could not be fit by MTPC, which requires at least 9 parameters to fit the 9 response rates. SDT and the 2HTM were fit to these data with the same parameters, but only 2 df. For similar reasons, Participant 13 was fit by SDT and the 2HTM with 1 df. Participant 8 did not produce enough datapoints to be fit by any model and was excluded. The summed values for each model include only those 8 participants who produced 10 response rates.

Table 8. Fit statistics for ratings and binary data, Experiment 2.

	G^2		AIC		BIC	
ID	SDT (12 df)	2HTM (10 df)	SDT	2HTM	SDT	2HTM
1	16.76	**8.12**	4874.4	**4869.8**	**4968.9**	4976.0
2	12.89	**11.55**	**5379.7**	5382.4	**5474.2**	5488.6
3	15.18	**11.37**	**4332.2**	4332.4	**4426.6**	4438.6
4	**53.91*****	63.72***	**3475.8**	3489.6	**3570.2**	3595.8
5	48.44***	**13.58**	5717.0	**5686.1**	5811.4	**5792.3**
6	33.05***	**28.29****	5768.9	**5768.1**	**5863.3**	5874.3
7	32.15**	**17.23**	5977.4	**5966.4**	**6071.8**	6072.7
8	17.26*	**11.44**	3792.0	**3790.2**	**3880.5**	3890.5
9	39.15***	**37.83*****	**4359.8**	4362.5	**4454.3**	4468.7
10	24.10*	**18.04**	5666.1	**5664.1**	**5760.5**	5770.3
11	47.69***	**25.90****	4837.0	**4819.2**	4931.4	**4925.4**
12	48.40***	**24.58****	5143.4	**5123.6**	5237.8	**5229.8**
13	31.53***	**28.43*****	**4774.5**	4775.4	**4868.9**	4881.6
14	**10.56**	12.90	**4688.2**	4694.5	**4782.6**	4800.7
Total	431.07***	**312.98*****	68786.4	**68724.4**	**70102.3**	70205.4
Group	82.80***	**46.00*****	85875.3	**85842.4**	86011.9	**85996.2**

**p<.05, **p<.01, ***p<.001.* The SDT and 2HT models were fit to Participants 3, 4, 9, and 12 with the same number of parameters as for the other cases but 11 and 9 df as these participants only produced 27 freely-varying response rates. For similar reasons, Participant 13 was fit with 10 and 8 df. Participant 8 was fit with 9 and 7 df and required only 15 (SDT) and 17 (2HTM) parameters.

Table 9. Best-fitting parameter values for Experiment 2, group data.

	SDT													
Condition	μ_o	σ_o	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	d_a		
Binary	0.76	1.27	1.21	1.02	0.77	0.58	0.39	0.30	0.22	0.01	-0.33	.67		
Binary (Binary Plus Ratings)	0.83	1.14	1.19	0.91	0.75	0.76	0.50	0.34	0.3	0.14	0.09	.77		
Ratings	0.74	1.26	1.24	0.69	0.47	-0.07	-0.75					.65		
Complete (Binary)	0.78	1.26	1.19	0.91	0.74	0.74	0.46	0.29	0.24	0.05	-0.02	.69		
Complete (Ratings)		1.25	0.71	0.48	-0.06	-0.74								
	2HTM													
Condition	p_o	p_n	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	s_l	s_m	s_h
Binary	0.30	0.19	0.14	0.19	0.27	0.35	0.44	0.48	0.51	0.61	0.72			
Binary (Binary Plus Ratings)	0.29	0.31	0.17	0.26	0.33	0.33	0.45	0.53	0.56	0.62	0.64			
Ratings (linear)	0.25	0.08	0.12	0.18	0.09	0.22	0.23	0.16						
Ratings (curved)	0.40	0.05	0.12	0.13	0.07	0.23	0.24	0.21				0.07	0.25	0.68
Complete (Binary)	0.40	0.07	0.12	0.19	0.24	0.24	0.34	0.42	0.44	0.53	0.56			
Complete (Ratings)			0.12	0.13	0.07	0.24	0.24	0.20				0.07	0.25	0.68

* d_a is not a free parameter in SDT, but is computed using the parameters μ_o and σ_o (see Introduction).

Table 10. Fit statistics for data from Experiment 2, using ratings ROCs from the unbiased (50% target) condition.

Ratings		Ratings and Binary					
	G^2	G^2		AIC		BIC	
ID	SDT	SDT	2HTM	SDT	2HTM	SDT	2HTM
1	7.19**	16.73	**11.11**	1890.7	**1889.0**	**1975.7**	1984.7
2	4.92	34.48***	**12.16**	2021.2	**2002.8**	2100.3	**2092.6**
3		**14.15**	21.91	**1670.8**	1682.6	**1749.9**	1772.3
4		**29.45**	64.22	**1509.6**	1548.4	**1582.8**	1632.2
5	3.05	19.21	**13.78**	2039.6	**2038.2**	**2124.6**	2133.8
6	**1.11**	**15.68**	25.48**	**1994.1**	2007.9	**2079.1**	2103.6
7	6.37	16.82	**15.42**	**1976.6**	1979.2	**2061.6**	2074.8
10	**1.13**	**11.56**	33.12***	**1895.4**	1921.0	**1974.5**	2010.7
11	5.70	47.76***	**28.92*** **	1920.3	**1905.5**	2005.3	**2001.1**
12	0.94	14.87	**9.42**	1756.2	**1754.7**	**1829.4**	1838.6
13	4.05*	30.51***	**29.96*** **	**1788.1**	1791.5	**1861.3**	1875.4
Total	34.46**	**251.22*** **	265.50***	**20462.6**	20520.9	**21344.5**	21519.7
Group	4.87	**34.21*** **	35.88***	**30594.5**	30600.2	**30722.1**	30743.7

**p<.05, **p<.01, ***p<.001.* Participants 3 and 4 were excluded from the ratings-only fits for producing fewer response rates than there were parameters to fit them. Participants 8, 9, and 14 were excluded from all of the fits as they produced fewer than 3 distinct points in the ratings data. Fewer than 10 freely-varying response rates were produced in the ratings data for the remaining participants, hence only SDT was fit to those data.

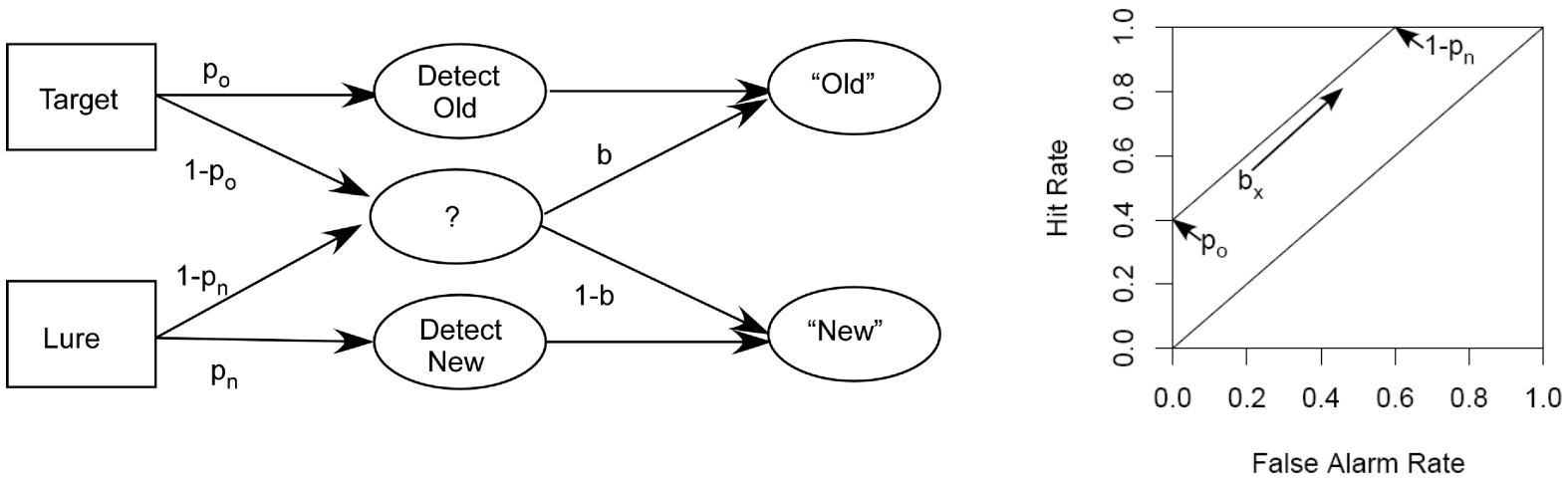

Figure 1. The double high-threshold model (left panel) and its implied binary receiver operating characteristic (right panel).

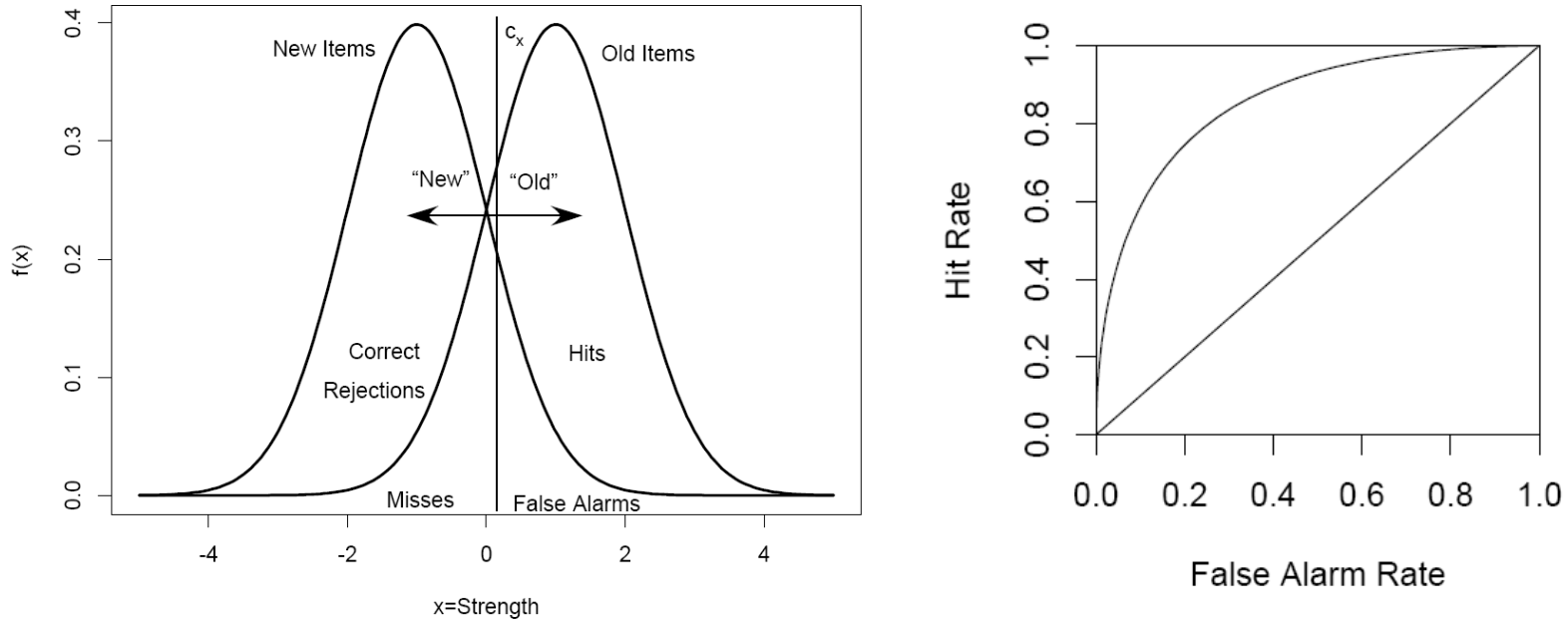

Figure 2. Signal detection theory (left panel) and its implied receiver operating characteristic (right panel).

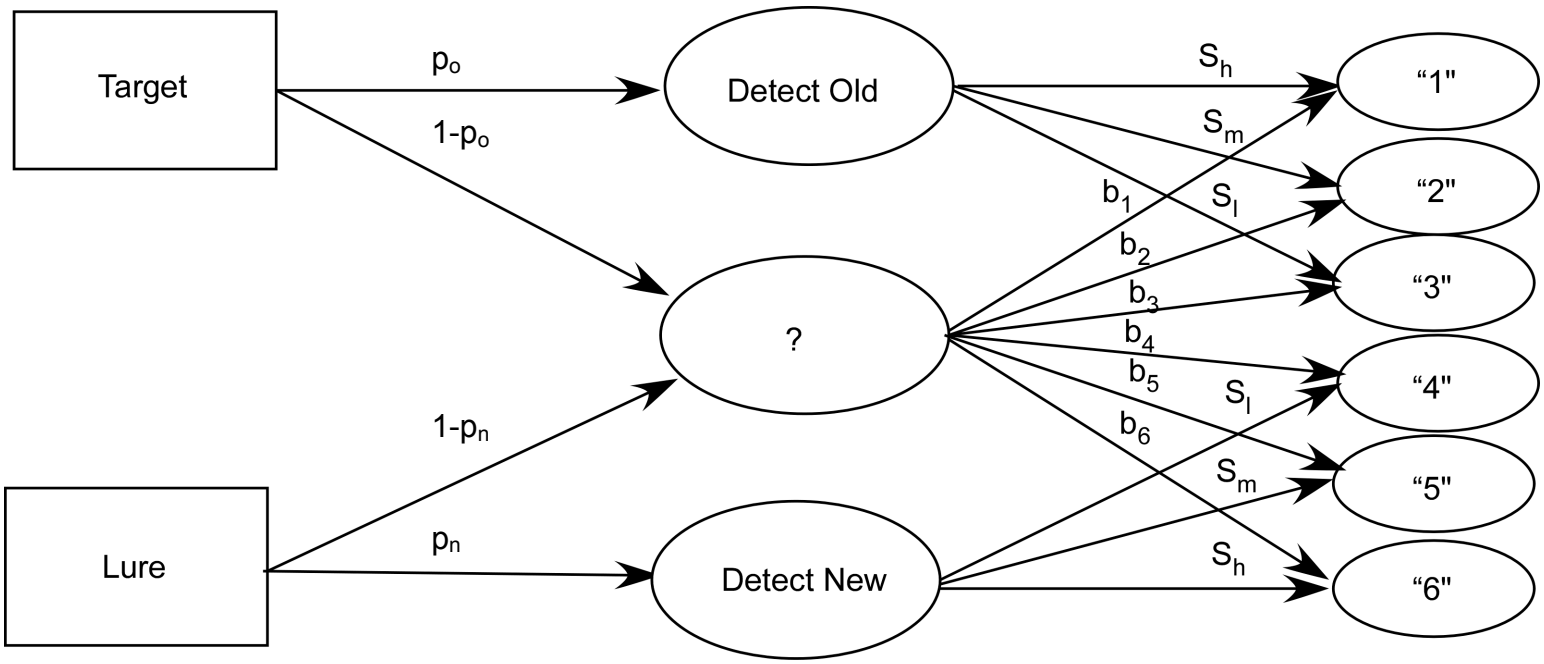

Figure 3. The multinomial processing tree model adapted from Klauer and Kellen (2010; 2011).

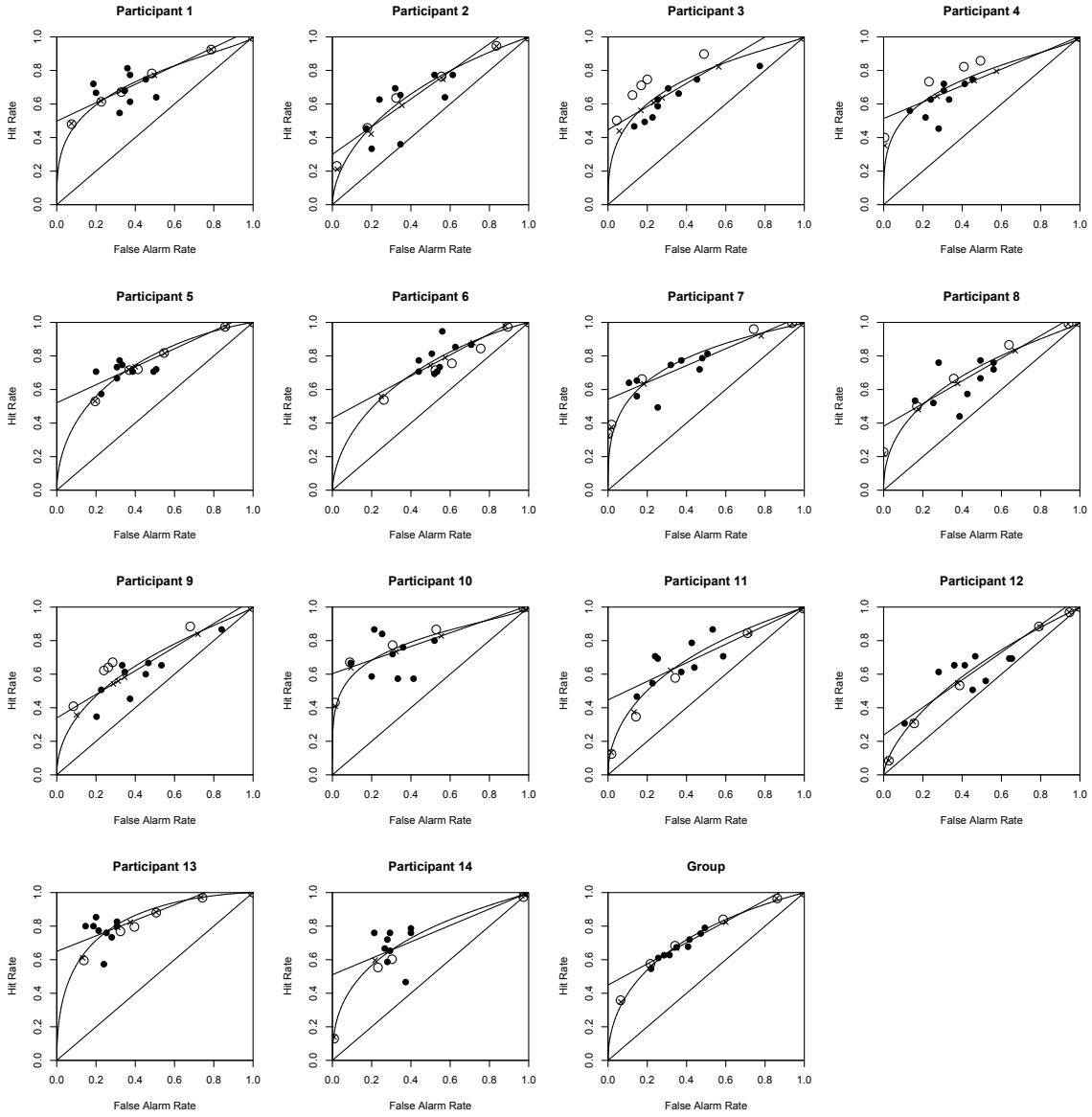

Figure 4. The ratings (open circles) and binary (filled circles) data of Experiment 1 fit by SDT and the 2HTM. SDT predictions for both functions are shown as a curved line and 2HTM predictions for the binary data are shown as a straight line. 2HTM predictions for the ratings data are shown as crosses.

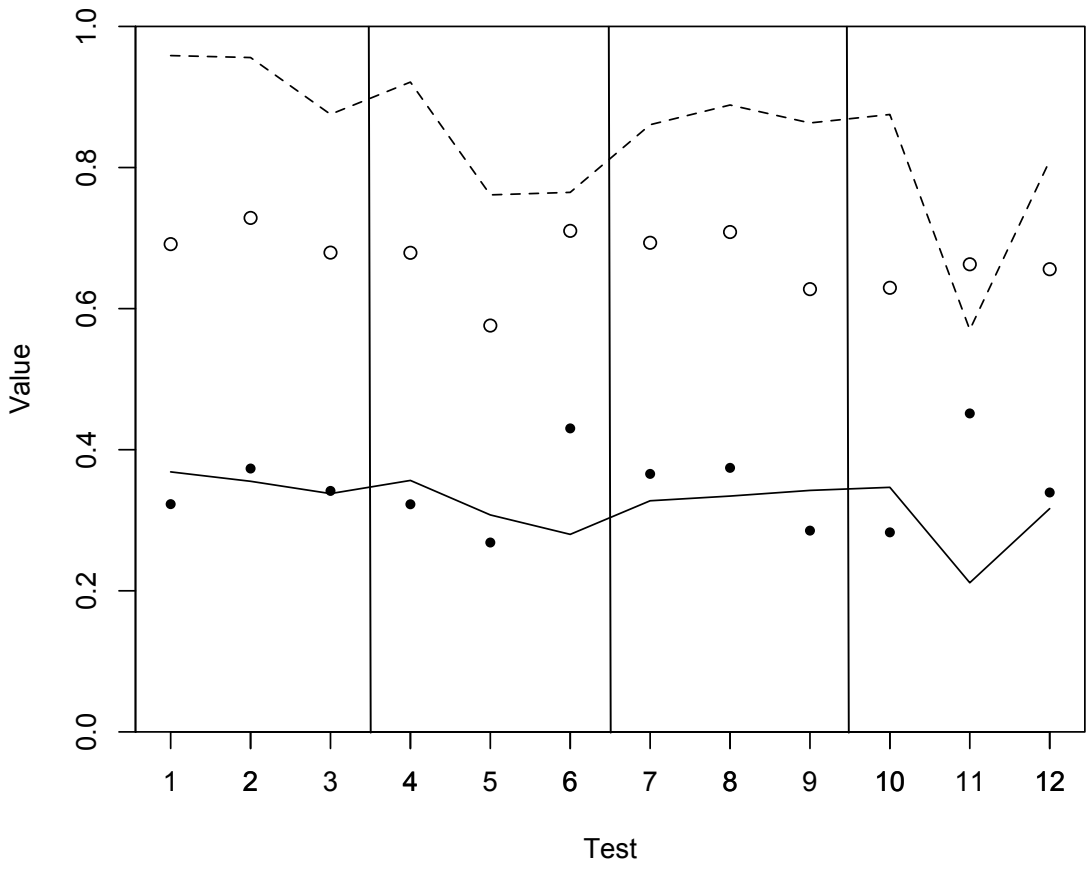

Figure 5. Hits (open circles), false alarms (filled circles), d_a (dotted line) and P_r (solid line) plotted as a function of test in Experiment 1, arranged chronologically. Vertical lines separate tests from separate sessions. d_a was calculated using the slope parameter obtained from the group fits, namely 1/1.23.

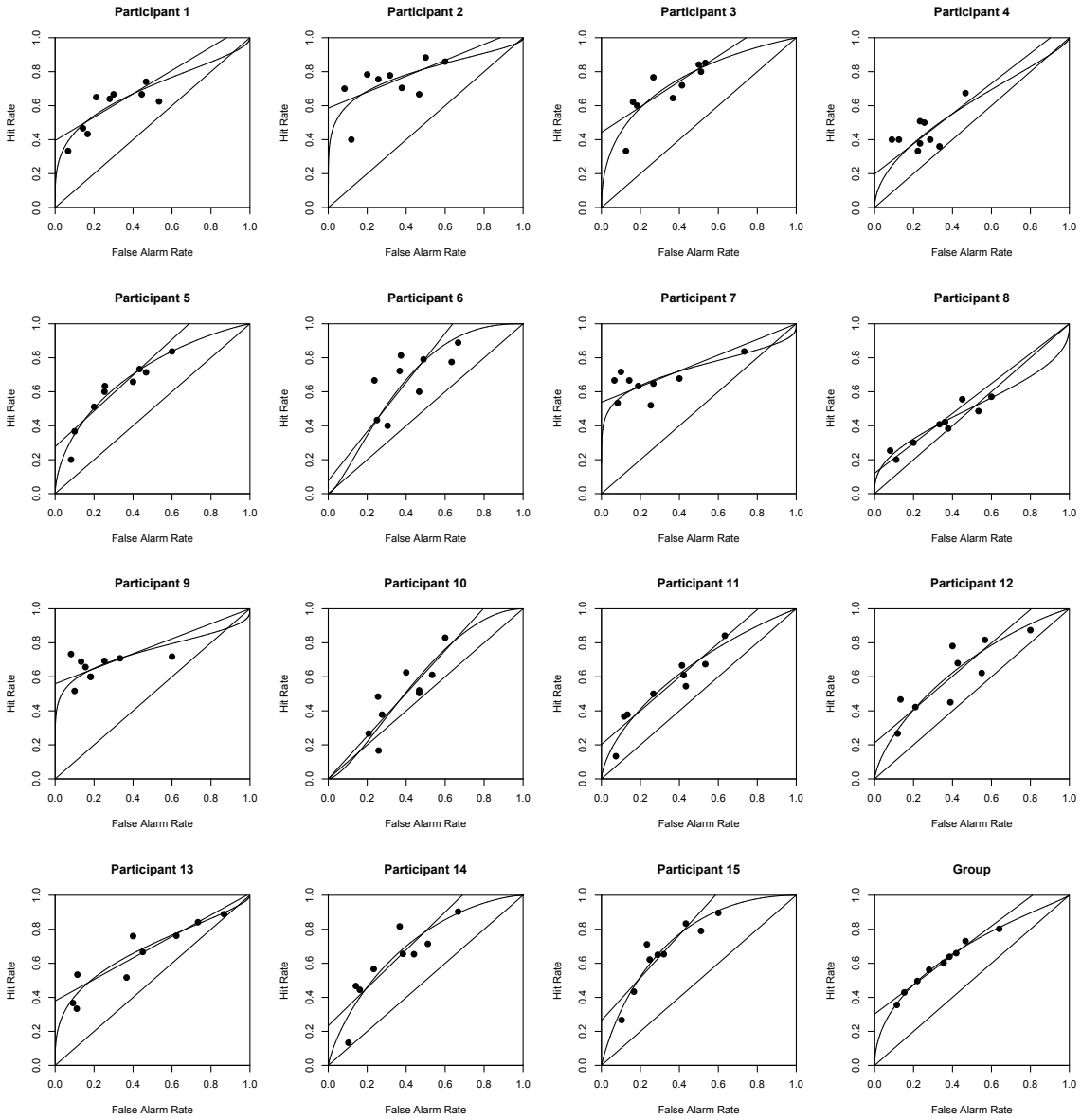

Figure 6. The binary only (filled circles) data of Experiment 2 fit by SDT and the 2HTM. SDT predictions are shown as a curved line and 2HTM predictions are shown as a straight line.

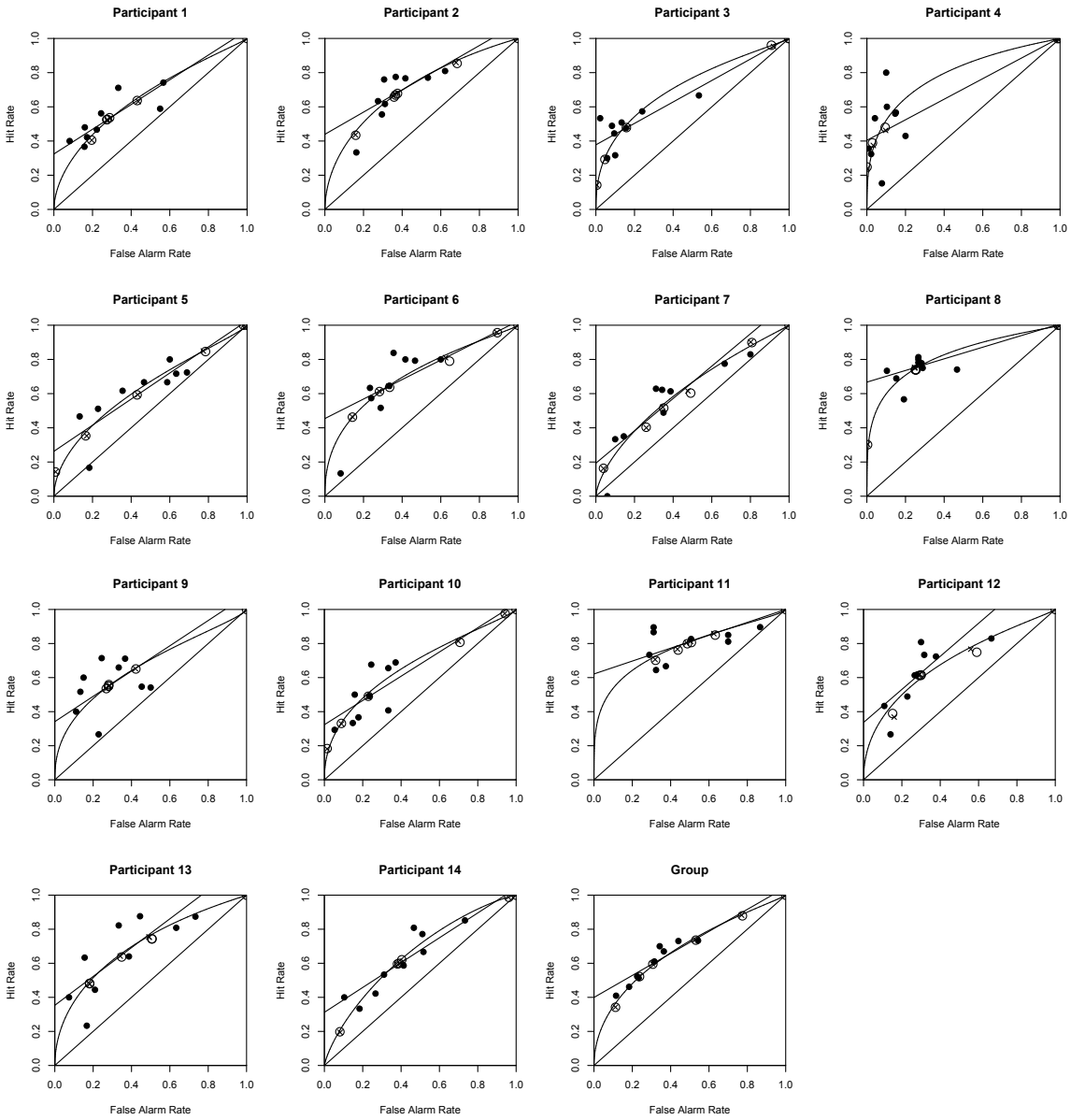

Figure 7. The ratings (open circles) and binary data from the same condition (filled circles) of Experiment 2 fit by SDT and the 2HTM. SDT predictions for both functions are shown as a curved line and 2HTM predictions for the binary data are shown as a straight line. 2HTM predictions for the ratings data are shown as crosses.

APPENDIX

MODEL EQUATIONS FOR EXPERIMENTS 1-2

2HTM

$H = p_o + (1-p_o)*b_x$

$M = (1-p_o)*(1-b_x)$

$F = (1-p_n)*b_x$

$CR = p_n + (1-p_n)*(1-b_x)$

*Where x corresponds to a particular bias condition.

2HTM, Ratings Data

Old items

$P(\text{“1”}) = p_o + (1-p_o)*b_1$

$P(\text{“2”}) = (1-p_o)*b_2$

$P(\text{“3”}) = (1-p_o)*b_3$

$P(\text{“4”}) = (1-p_o)*b_4$

$P(\text{“5”}) = (1-p_o)*b_5$

$P(\text{“6”}) = (1-p_o)*b_6$

New items

$P(\text{“1”}) = (1-p_n)*b_1$

$P(\text{“2”}) = (1-p_n)*b_2$

$P(\text{“3”}) = (1-p_n)*b_3$

$P(\text{“4”}) = (1-p_n)*b_4$

$P(\text{“5”}) = (1-p_n)*b_5$

$P(\text{“6”}) = p_n + (1-p_n)*b_6$

*Note that b_6 is not a free parameter, but is equal to one minus the sum of b_1 to b_5.

MPTC

Old items

$P(\text{“1”}) = p_o*s_h + (1-p_o)*b_1$

$P(\text{“2”}) = p_o*s_m + (1-p_o)*b_2$

$P(\text{“3”}) = p_o*s_l + (1-p_o)*b_3$

$P(\text{“4”}) = (1-p_o)*b_4$

$P(\text{“5”}) = (1-p_o)*b_5$

$P(\text{“6”}) = (1-p_o)*b_6$

New items

$P(\text{“1”}) = (1-p_n)*b_1$

$P(\text{“2”}) = (1-p_n)*b_2$

$P(\text{“3”}) = (1-p_n)*b_3$

$P(\text{“4”}) = p_n*s_l + (1-p_n)*b_4$

$P(\text{“5”}) = p_n*s_m + (1-p_n)*b_5$

$P(\text{“6”}) = p_n*s_h + (1-p_n)*b_6$

*Note that s_l is not a free parameter, but is equal to $1-(s_h + s_m)$.

SDT, Binary Data

$H = \Phi[(\mu_o-c_x)/\sigma_o]$

$M = 1-H$

$F = \Phi(-c_x)$

$CR = 1-F$

*Where x corresponds to a particular bias condition. The function $\Phi(z)$ returns a value P(z) from the inverse standard normal cumulative distribution function for a given value z.

SDT, Ratings Data

Old items

$P(\text{“1”}) = \Phi[(\mu_o-c_1)/\sigma_o]$

$P(\text{“2”}) = \Phi[(\mu_o-c_2)/\sigma_o] - \Phi[(\mu_o-c_1)/\sigma_o]$

$P(\text{“3”}) = \Phi[(\mu_o-c_3)/\sigma_o] - \Phi[(\mu_o-c_2)/\sigma_o]$

$P(\text{“4”}) = \Phi[(\mu_o-c_4)/\sigma_o] - \Phi[(\mu_o-c_3)/\sigma_o]$

$P(\text{“5”}) = \Phi[(\mu_o-c_5)/\sigma_o] - \Phi[(\mu_o-c_4)/\sigma_o]$

$P(\text{“6”}) = \Phi[(c_5-\mu_o)/\sigma_o]$

New items

$P(\text{“1”}) = \Phi(-c_1)$

$P(\text{“2”}) = \Phi(-c_2) - \Phi(-c_1)$

$P(\text{“3”}) = \Phi(-c_3) - \Phi(-c_2)$

$P(\text{“4”}) = \Phi(-c_4) - \Phi(-c_3)$

$P(\text{“5”}) = \Phi(-c_5) - \Phi(-c_4)$

$P(\text{“6”}) = \Phi(c_5)$

REFERENCES

Akaike H. (1973). Information theory as an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Akademiai Kiado, Budapest.

Balakrishnan, J.D. (1999). Decision processes in discrimination: Fundamental misconceptions of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189-1206.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548-564.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 197-215.

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Zeitschrift Für Psychologie/Journal of Psychology, 215*(1), 52-60.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear – or are they? On premature arguments against the two-high-threshold model of recognition**.** *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 587-606.

Burnham, K.P., & Anderson, D.R. (2005). *Model selection and multi-model inference* (2nd Ed.). New York, NY: Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ, US: Erlbaum.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33A,* 497-505.

Dube, C., Rotello, C.M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*, 831-863.

Dube, C., Rotello, C.M., & Heit, E. (2011). The belief bias effect is aptly named: A reply to Klauer and Kellen (2011). *Psychological Review, 118,* 155-163.

Egan, J.P., (1958). Recognition Memory and the Operating Characteristic. USAF Operational Applications Laboratory Technical Note, No. 58-51, 1958. pp. ii, 32.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134-140.

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295-306.

Friedman, M. P., Carterette, E. C., Nakatani, L., & Ahumada, A. (1968). Comparisons of some learning models for response bias in signal detection. *Perception & Psychophysics, 3,* 5-11.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 500-513.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition, 36*(4), 703-715.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15*(5), 889-905.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1210-1230.

Heit, E., & Rotello, C.M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L. T. (2002). Regularities of source recognition: ROC analysis. *Journal of Experimental Psychology: General, 131*(4), 494-510.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 302.

Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition, 29*(8), 1176-1184.

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35*(4), 679-688.

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101*(1), 166-171.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review, 17*, 465-478.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (in press). *Psychological Review, 118,* 164-173.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*, 852-884.

Krantz, D.H. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308-324.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Macmillan, N.A., & Creelman, C.D. (2005). *Detection theory: A user's guide (2nd ed.).* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics, 66*, 406–421.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 380-387.

Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception & Psychophysics, 2*, 91-97.

McGeoch, J.A., & Irion, A.L. (1952). *The Psychology of Human Learning.* New York: Longmans, Green.

Mueller, S.T., & Weidemann, C.T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*, 465-494.

Onyper, S. V., Zhang, Y. and Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General*, *139*, 341-364.

Paivio, A. (1971). *Imagery and verbal processes.* Oxford, England: Holt, Rinehart & Winston.

Pratte, M. S., Rouder, J. N., Morey, R. D. (2010). Separating mnemonic processes from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 224-232.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 763-785.

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518-535.

Rotello, C.M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1317-1330.

Rotello, C. M., Macmillan, N. A., & Hautus, M. J. (under review). Source and item recognition memory: Comparing likelihood-ratio signal-detection, dual-process signal-detection, and multinomial processing models. *Manuscript submitted for publication.*

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70*, 389-401.

Schulman, A. I., & Greenberg, G. Z. (1970). Operating characteristics and a priori probability of the signal. *Perception & Psychophysics, 8*, 317-320.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Sherman, J.W. (2006). Clearing up some misconceptions about the Quad Model. *Psychological Inquiry, 17,* 269-276.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33*(1), 151-170.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1499-1517.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34-50.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379-1396.

Swets, J.A. (1986a). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100-117.

Swets, J.A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks : Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181-198.

Treisman, M., & Faulkner, A. (1984). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology, 37*, 199-215.

Underwood, B. J. (1957). Interference and Forgetting. *Psychological Review, 64,* 49-60.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26*, 582-600.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*(2), 254-262.

Wagenmakers, E-J., & Farrell, S. (2004). AIC model selection using Aikaike weights. *Psychonomic Bulletin and Review, 11,* 192-196.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152-176.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characterstics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1415-1434.